Lexical Semantic Similarity and its Application to Business Catalog Retrieval

by

Jian (Jay) Jiang

A thesis

presented to the University of Waterloo

in fulfilment of the

thesis requirement for the degree of

Doctor of Philosophy

in

Management Sciences

Waterloo, Ontario, Canada, 1998

0-612-32834-1

The University of Waterloo requires the signatures of all persons using or photocopying this thesis. Please sign below, and give address and date.

Abstract

This thesis targets the problems of language variability (i.e. polysemy and synonymy) from the viewpoint of lexical semantic similarity — a measure of semantic/conceptual similarity between pairs of lexicalized concepts represented in words or terms. As is often the case for many tasks in information retrieval (IR) and natural language processing (NLP), a job is decomposed to the requirement of resolving the semantic relation between lower-level constituents such as words or concepts. One needs to develop a consistent, widely applicable computational model to assess this type of relation.

We believe that a proper identification of similarity between concepts would contribute significantly in resolving semantic ambiguity in general. We start by looking at the fundamentals of the concept of similarity, its assumptions and characteristics. A new framework of universal object comparison and similarity determination scheme is then constructed in set-theoretic notions. This is in response to the observation that there is generally a lack of systematic classification and definition of various similarity formulae. Typically, a similarity formula or metric is directly employed in a problem without much theoretical justification and the assumptions behind it are not stated explicitly. In our study, rather than directly stipulate a similarity definition, we intend to derive it from a set of reasonable and intuitively justifiable assumptions. We then argue that this framework provides a general account for modeling object comparisons, and some of the specific comparison schemes can be further abstracted and quantified using information-theoretic notions so that a simple computational means of measuring universal object similarity can be achieved.

To realize such a computational object comparison scheme, we propose a new model of measuring lexical semantic similarity given the context of a lexical taxonomy. This model enhances the graph distance approach by properly quantifying the weight of each edge along the shortest path that links two concept nodes in the taxonomic hierarchy.

The lexical semantic information derived from taxonomy structure essentially secures a solution to determining the 'commonality' of two objects in information-theoretic fashion, which is crucial to a realization of a computational means of resolving universal lexical semantic similarity. This also allows us to develop a unified view of various similarity measures based on taxonomic knowledge, given the background of our general framework about object comparison schemes. Contrary to the common view that object 'commonality' dictates similarity, both our theoretic model and later empirical verifications have demonstrated that object 'difference' is perhaps a better approximation to the similarity measure when object content is measured by its information content.

This core similarity model is then applied to several levels of applications in NLP and IR. First, a word-pair similarity ranking experiment was conducted. The results indicate the proposed similarity measure ('difference' approach), compared with other related computational models, achieves a result that is closest to humans' performance when the same task was replicated. Second, a simple word sense disambiguation algorithm is

developed based on the local contextual information obtained from words surrounding the target ambiguous word. The empirical evidence from the test of tagging all nouns in a running text verifies that the proposed similarity model can generate better performance than other related similarity models in this 'intermediate'-level NLP application.

To raise the complexity of the issue of semantic similarity we move from conducting a single, elemental concept pair similarity comparison to a multi-layered compound concept (phrase-like) similarity comparison. A final and more practical application of the proposed similarity model is in the areas of text retrieval and document classification. Originating from a project to develop an Electronic Industrial Directory (EID) system, a prototype business catalog retrieval system is designed and implemented. The main function is to locate the relevant catalog headings under which a product/service description would belong.

In order to provide an appropriate context for lexical-level similarity comparisons, a shallow parsing algorithm is designed to capture both syntactic and semantic information in both catalog headings and queries. The weak technologies employed here require no pre-existing domain specific knowledge structures. Hence the resultant model has an appeal to wider domains of application. We developed various methods in both decomposing complex linguistic constructs into single lexicalized items wherein the developed lexical similarity methods can be applied, and aggregating such calculated subcomponent similarities to an overall similarity determination. For the within-subcomponent parsing and similarity determinations, we designed several algorithms to tackle the syntactic ambiguity problems such as compound nouns and complex phrases analysis. For the between-subcomponent similarity determinations, we constructed a framework and proposed two dynamic subcomponent weighting schemes in terms of subcomponent information content value. A prototype system was developed and evaluated against the benchmark of the classical vector space model. The analysis of results indicates that there is a significant improvement over the benchmark (both precision and recall are increased by about 10%).

Acknowledgements

I am very grateful to my advisor, Dave Conrath, for his support, for his encouragement, and for his guidance. He has provided just the right mix of motivation, feedback, friendship, and patience, and for this I will appreciate deeply for the rest of my life.

I would like to thank the members of my dissertation committee: Graeme Hirst, Ji-Ye Mao, John Moore, and Frank Tompa. They are individually extraordinary, and together they form a committee with enormous intellect and personality.

I want particularly to thank Graeme Hirst and Frank Tompa for their deep insight in this multidisciplinary research area, their constructive criticism and valuable feedback, which have significantly contributed to the quality of the finished work.

Special thanks are deserved by Dekang Lin, Alexander Budanitsky, Chrysanne DiMarco, Reem Al-Halimi, Peter Vanderheyden, Vlado Keselj, Toby Donaldson, Stuart Gill, Philip Edmonds, Liyang Yu, and Norm Archer, with whom I have had an important discussion or a shared insight over time.

And now my deepest gratitude and love go to my wife Xia Man, without whom this dissertation and research behind it would hardly ever been realized. There is nothing I can ever do to repay the encouragement, sacrifice, and patience she provided.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Setting

This thesis is about the study of word semantic similarity. It is about developing computational models for computers to better 'recognize' the relationship between elemental concepts represented in human language, and applying the resultant models to various natural language processing tasks such as lexical semantic disambiguation and conceptual information retrieval. Just as for other work in natural language processing, the ultimate goal is to have computers 'understand' language in the sense that humans do; and for this specific pursuit, it is expected that computers can simulate this human cognitive process (determining word semantic similarity) to produce results that are close to humans' judgments.

Fundamental to all sciences are observations and determinations of similarity between phenomena. Perceptual similarity, the process of observing similarity from stimuli via the sensory systems, is one of the very first steps in human cognitive processes. Conceptual similarity is then a deeper understanding of the relationship between phenomena that are abstracted in humans' minds. Similarity has been used as a concept since the days of Aristotle to bring order from chaos, and to construct structure from variety. Many social sciences, and particularly psychology, consider it as one of the fundamental concepts. In many natural sciences, such as biology and ecology, determining similarity between objects is a prerequisite to many other developments such as categorization and classification.

The demand for such treatment can also be observed from modern disciplines like computer science and in particular its subdiscipline, artificial intelligence. The problems of *language variability*, the characteristics of polysemy and synonymy that exist in words of natural language, have always been a challenge in the fields of Natural Language Processing (NLP) and Information Retrieval (IR). In many cases, humans have little difficulty in determining the intended meaning of an ambiguous word, while it is extremely difficult to replicate this process computationally. For many tasks in psycholinguistics and NLP, a job is often decomposed to the requirement of resolving the semantic relations between words or concepts. One needs to come up with a consistent computational model to assess this type of relation. From a practical standpoint, a solution is expected to be easily implemented, less resource consuming as compared to a full-scale language understanding, and have a wide domain of applicability.

Although the study of similarity has a long history, most often the representation and measuring of similarity are highly tied to a specific domain of study. Up to now, there is no universal definition of similarity. Many classical quantitative definitions have been used under various scenarios. Often they are a prerequisite or meta-model in other analysis models. For example, cluster analysis has its goal to aggregate objects on the basis of object similarities. Pattern recognition is identifying the similarity between the new pattern and the existing ones. Typical similarity methods are assimilated and integrated into other larger, more sophisticated models. For example, in the classic statistical approach to IR, the cosine retrieval function is a similarity measurement between the query and a specific document. Moreover, although many similarity formulae were validated through empirical investigations, the underlying assumptions behind them are not stated explicitly, and hence they are subject to rigorous theoretical questioning and arguments. Bearing this in mind, we start by looking at the fundamentals of the concept of universal object comparison and similarity determination, their assumptions and characteristics. Rather than directly stipulate a similarity definition, we intend to derive it from a set of reasonable and intuitively justifiable assumptions. The emphasis is on the abstraction of the domain, so that the derived formula has more universal applicability.

We then further explore the relationship between similarity and its related concepts. We demonstrate how these object comparison schemes can be quantified and applied to a wider domain, and in particular, those domains that have a probabilistic model. This brings us closer to the treatment of the bottleneck problems in NLP and IR that have been identified. Equipped with distributional information from corpus data, and linguistic knowledge from a manually built taxonomy, we are able to develop a proper computational method for determining semantic similarity between words and concepts.

This core similarity model is then applied to other levels of applications in NLP and IR. A simple word sense disambiguation (WSD) algorithm is developed which uses the local contextual information obtained from words surrounding the target ambiguous word. The main purpose for performing this WSD task is to demonstrate the effectiveness of such a simple method, and particularly to seek empirical evidence which would further support the claim that the proposed similarity model can generate better performance than other related similarity models in NLP applications.

To raise the complexity of the issue of semantic similarity we move from conducting a single, elemental concept pair similarity comparison to a multi-layered compound concept (phrase like) similarity comparison. A final and more practical application of the proposed similarity model is in the areas of text retrieval and document classification. This actually was the motivation of the research work that was targeting at building an Electronic Industrial Directory (EID) — a Yellow Pages-like online business directory for storing and retrieving trade related business information. The main function of the EID is to match the product/service descriptions provided by a purchaser with that of a supplier so that a potential business exchange can be further pursued. One of the decomposed tasks of the EID functions is to organize the business directory in a fashion that each company product/service description is properly classified under a particular business category. The Standard Industrial Classification (SIC), which covers a wide range of business category information, is chosen to represent the framework of the EID database. Therefore, a function is needed, for both the business classification and subsequent query retrieval, to locate the relevant SIC heading(s) under which a product/service description would belong.

Both users' queries and SIC headings are often more complex than a single atomized concept. To accomplish the task for the SIC headings matching and retrieval, we need to develop algorithms to parse both SIC headings and the incoming query, to compare their corresponding decomposed parts, and to arrive at a weighted overall semantic similarity value between them. The semantic similarity scheme developed for comparing single atomized concepts can then be applied to the decomposed processes.

The underlying premise of the work is that conceptual matching between the query and a document is the ultimate solution to an information retrieval task. Therefore an approach that directly applies semantic matching schemes would be a significant step towards realizing this goal. Moreover, in developing such a method, advantage can be taken by combining both statistical information from unstructured data (e.g. corpora) and linguistic knowledge from highly structured and organized constructs (e.g. lexicons, thesauri). The achievement of this work confirms the benefit of the growing trend of integrating linguistic knowledge into corpus statistics in the area of natural language processing (NLP) and information retrieval (IR).

1.2 Chapter Summaries

Chapter 2. We first discuss some general background and theories about similarity and its related concepts. Then we present an extensive review of four major similarity models with regard to various types of data to be compared in the similarity study.

Chapter 3. We go back to the single and fundamental issue about similarity – its definition. As we have observed, most similarity formulae are limited to a certain domain and in particular lack a solid theoretical justification. We therefore seek, under a more generic scenario, a quantitative definition that is both theoretically sound and widely applicable. Rather than directly stipulating a similarity definition, we apply set theory to model universal object comparison and derive a similarity theory that is based on a set of reasonable and intuitively appealing assumptions. The derived result can be seen as a generalization of some of the classical definitions of similarity. We also discuss other similarity related concepts and the relationships among them. A computational model is proposed for the developed object comparison scheme in information-theoretic terms, which become the basis for theory development in the remaining parts of the thesis. Finally, we present an informal argument that one of our defined similarity related concepts can be regarded as an instantiation of one of the similarity models reviewed in Chapter 2.

Chapter 4. We apply the information-theoretic definition of similarity derived in the last chapter to a very basic scenario — measuring semantic similarity between lexicalized concepts represented by words or terms. We first discuss the concept of semantic similarity and its more general concept — semantic association. We then briefly review two main approaches to semantic similarity using corpus statistics or taxonomic knowledge. After introducing the recently developed and broad-coverage lexical taxonomy WordNet, we describe the development in modeling semantic similarity between concepts in a taxonomy. Following that we present a new similarity measure which is an integration of previous methods. We explain how the proposed semantic relevance model fits into the similarity theory that we developed in Chapter 3. In section 4.8, experiments are conducted to evaluate various computational similarity models by comparing them against human judgment on ranking similarity between word pairs.

Chapter 5. In this chapter, we investigate an application of the semantic similarity model proposed in Chapter 4 to an 'intermediate' level NLP task: Word Sense Disambiguation (WSD). We employ a simple approach to WSD without the need of any previously trained (tagged) data. Like most WSD methods, we intend to utilize the cues from the information of the target word's contextual content. An experiment is performed to compare the proposed similarity models with other computational models that are used in the experiment of chapter 4. The results further verify the improvement of the proposed model over other models.

Chapter 6. We describe the implementation details and the performance evaluation of a prototype SIC headings search and retrieval system. In order to provide an appropriate context for lexical level similarity comparison, a shallow parsing algorithm is designed to capture both syntactic and semantic information in the headings and queries. The weak technologies employed here require no pre-existing domain specific knowledge structures. Hence the resultant model has appeal to a wider domain of application. In testing the performance of the implemented prototype, we import sample queries from real world business trade inquiries. Several batches of experiments are conducted to compare the effectiveness of various decision factors in the retrieval model. In the last section we discuss the significance and applicability of applying the SIC headings search model to other similar tasks.

Chapter 7. We summarize the contributions of the thesis, and present some thoughts on future work.

Chapter 2

Background on Similarity

2.1 Introduction

Similarity is a widely used concept in many disciplines, such as psychology, cognitive science, biology, microbiology, and computer science. Studies vary from using empirical verification to theoretical modeling. A large number of mathematical formulae have been contrived to define it in order to capture the essence and relationships between phenomena as well as reflecting the human mental judgment of such relationships.

In this chapter, we first discuss the background and characteristics about object, concept, and similarity and its related concepts. Then we present four major similarity models ranging from the well-known classical accounts to some newly developed frameworks. Accompanying the classical models, a list of various representative similarity definitions and mathematical formulae is given with regard to the types of data to be compared in the similarity study. Finally, we conclude this chapter with a summarization of the models presented and a discussion of their implications for the development of proposed models for this thesis.

2.2 Object, Concept, and Similarity

In the behavioral sciences, the process of observing similarity is typically referred to as a cognitive process when people react to external stimuli. This happens when people make comparisons of objects (space dimension), retrieve stored memories (time dimension), etc. Observing similarity appears also to have an important role to play in learning, reasoning, inference, and problem solving (Watanabe 1985). Before we explore the similarity issue, we first discuss the compositional elements under the concept of similarity, in particular, what constitutes the objects to be compared in a human's mind. We then give a general account of the characteristics of similarity and its related concepts.

2.2.1 Object and Concept

For humans an *object* in the universe has several layers of meaning. First, it is the actual object itself, i.e the single, physical and unique existence, independent of any human being's perception or description. Second, it is the mapping of such an object into a human's mind, i.e. the *concept* of the object. A concept is a mental construct whereby a person abstracts the physical object or some other relations into his/her mind. It is assumed that our minds can 'contain' knowledge of the world that will interpret the stimulus from the real world and then construct a unique mental representation. In this scenario, we are focusing on a collective generation of a concept that refers to some abstract societal norm, which is more or less the same in most individuals' minds. Third, it is a *representation* of the concept of the object as perceived by humans. This is a mapping of the original object from its mental depiction (concept) back to the physical world in order for people to communicate. This is typically in a form of words of a language, a picture or sound, or in other media that convey the essentials of the *concept* of the object. Note that due to the variety, flexibility and complexity of media and human communication channels, this process could be a one-to-many relation. A schematic notion of the process account for identifying and expressing an object can be depicted in Figure 2.1:

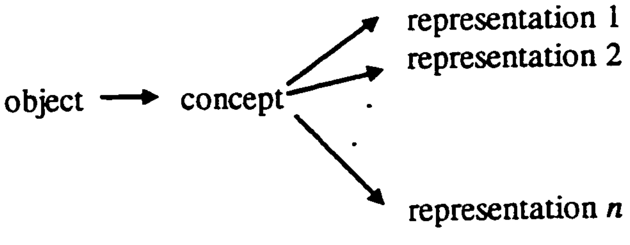


Figure 2.1 A process-oriented cognitive view about an object

Figure 2.1 provides a process account on how objects/phenomena are viewed, reflected and represented. Another similar view is the famous 'meaning triangle' model which emphasizes the reverse flow of meaning representation, and in particular the relation among each mapping process (Ogden and Richards 1923). We present it here in Figure 2.2. The dotted baseline of the triangle indicates an indirect relation between Symbol and Referent. In fact, this is the source of almost all the complexities and difficulties encountered in areas where languages and symbols are involved.

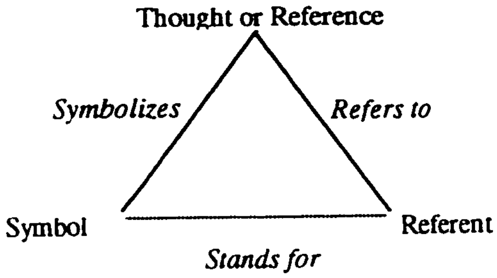


Figure 2.2 Meaning triangle by Ogden and Richards

For the study of an object, it would make no sense to study directly the first sense of objects. In most cases, it is impractical and even impossible to manipulate the physical objects due to space, time and other constraints. It would also be very difficult to study the second sense since the pure concept of an object exists only in one's mind. Most likely, the object that people directly observe and study is in its third sense — an abstraction that transforms the formless concept into a materialized form exhibited in the physical world. Therefore, when people talk about objects and concepts, they are expressing them in one of their representations. In the remaining parts of the discussion, unless specified explicitly, the object or concept we refer to is the last sense of the meanings. Since this is the result of a possible one-to-many mapping, it is anticipated that the complexity of studying it would be high[1].

[1] This is one of the explanations of the problems of language variability we identified earlier, as language is one of the (main) modes for humans to represent concepts.

Henceforth, when we discuss similarity among objects, we actually mean similarity among their concepts. We adopt the classical approach taken to study concepts, in which a *concept* is the mental notion by which an intelligent being is able to understand some aspect of the world (Murphy and Medin 1985). A concept is described by a set of attributes with associated values that define its intensional properties, and a set of primitive objects that clarify its extensional boundaries. For our purposes, the intensional properties of concepts constitute their representations, which we will manipulate and analyze to evaluate similarity.

Using Formal Concept Analysis (Priss 1996), a theory for conceptual data structuring, we can present a formal definition of a concept. We start with the notion of *formal context*. This is defined as a triple (O, A, R), where O is the set of *formal objects*, A is the set of *formal attributes*, and R is a binary relation mapping O and A. The notion oRa denotes that the object o has the attribute a. For a given subset of objects $X \subseteq O$, its prime, X', yields all the common attributes of that set: $X' := \{a \in A \mid oRa \text{ for all } o \in X\}$. Similarly, for a given subset of attributes $Y \subseteq A$, its prime, Y', yields all common objects of that set: $Y' := \{o \in O \mid oRa \text{ for all } a \in Y\}$. A pair (X, Y) is said to be a *formal concept* of the context triple (O, A, R) if $X \subseteq O, Y \subseteq A, X = Y'$, and $Y = X'$. X is then called the *extent*, and Y the *intent*. Notice that the intensional and extensional definition are dual/interdefinable. That is the intension is composed of just those attributes that are true of all the extensional members, while the extension is composed of just those objects that possess all the concept's intensional attributes.

2.2.2 Similarity and Related Concepts

Once we clarify the component parts of similarity, we can explore it as a whole. *Similarity*, in essence, is the result of the process of exploring and deriving an abstract relationship between two or more objects. There are many relationships between objects from different perspectives. One of them is the *affinity* or *proximity* relation, which is a more general term for 'similarity'. It includes distance, dissimilarity, substitutability, association, correlation, confusion, etc. Among them, similarity is the most widely employed and studied concept.

There are two basic views of similarity (Feger and Boeck 1993). On the one hand, similarity is the result of a subjective experience when reacting to the stimulus. The resulting observations are called *similarity judgments*. On the other hand, similarity may be derived, usually from descriptions of objects and their components (e.g. attributes). The former represents the direct empirical study, while the latter represents an indirect calculation approach. In this particular research, the latter view is pursued for our computational approach.

No matter whether similarity is the result of direct observations or an indirect derivation from the object's characteristics, the final representation of it, s_{ij}, for objects i and j, is usually quantified as a numerical value (Gregson 1975:16). Traditionally, it is within an interval of [0, 1]. Hence, $1-s_{ij}$ defines *dissimilarity*. Often, certain assumptions are presupposed in a classical similarity model:

- Maximum self-similarity. $s_{ii} > s_{ij}$ for $i \neq j$.
- Symmetry. $s_{ij} = s_{ji}$.
- Monotonicity. The similarity will increase when there is a common part added to the two compared objects.

Maximum self-similarity implies that an object is most similar to itself. In many cases, the result of a similarity comparison should be symmetrical. Monotonicity reflects a characteristic in the process of object comparison.

A concept very closely related to similarity is *distance* or *metric*: d_{ij}. Originated from a concept in geometry, distance has been introduced into many disciplines in addition to the natural sciences. A distance/metric satisfies four axioms:

I Non-Degeneracy: $d_{ij} = 0$ if and only if $i = j$

II. Triangular Inequality: $d_{ij} \le d_{ik} + d_{kj}$

III. Non-Negativity: $d_{ij} \ge 0$

IV. Symmetry: $d_{ij} = d_{ji}$

As we will see in later discussions, many mathematical similarity formulas meet these criteria and hence are a kind of metric. However, empirical studies have found that there are occasions that violations of some of the metric axioms (as well as the similarity assumptions) do exist in humans' similarity judgments. More sophisticated models have been constructed with the capability of dealing with various complex scenarios.

2.3 Similarity Models

We now turn to describe some of the major models of similarity that have been introduced in both natural science and social science studies. The classification scheme follows the notion from Goldstone (1998) where four major psychological similarity models are identified: geometric/spatial, feature-based, alignment-based, and transformational. The first two are the dominant models in classical similarity studies and will be discussed in detail. The last two will be sketched as their development is still at its early stage and lacks concrete operational vehicles[2]. They nevertheless exhibit a deeper understanding of the subject and indicate the direction of the development of similarity study.

Each of the first two classic models will be further decomposed into specific treatments according to the various types of data. Normally, data types can be divided into two major categories: qualitative and quantitative. Under the qualitative notion, there are further subcategories such as: dichotomous/binary/logical types, nominal/multilevel types, and ordinal types. For the quantitative category, there are numerical or interval types.

[2] In fact, the later development and applications of our similarity models can be regarded as a manifestation of both approaches.

2.3.1 Geometric/Spatial Models

Among the earliest and most influential approaches to similarity are the geometric/spatial models. They represent objects as real-valued vectors and specify a distance metric defined on vector differences or scalar products. The best known of these are multidimensional scaling (MDS) and factor analysis. The ***multidimensional scaling models*** (Shepard 1962) assume that objects are decomposed into values along component dimensions (e.g. attributes). Each object is then represented as a point in a multidimensional space. The similarity between objects is a decreasing function of the distance between their values on these dimensions.

If n is the cardinality of the object set O and m the cardinality of the attribute set A, then we may consider the objects to be n points in a high dimensional space of $\Re^m$, or, alternatively, attributes to be m points in $\Re^n$. For the former case, each object can be represented as a vector:

$$\mathbf{x}_i = \{x_{ij} \mid j = 1,\ldots,m\},$$

where each of the m entries can have different data types for a specific attribute value.

2.3.1.1 Quantitative Data

Geometric models are very suitable in calculating the similarity for quantitative data.

Basic MDS Model

Given a multidimensional space where each object is represented as a point in the space, the similarity between two points is taken to be inversely related to their distance, which is computed by the Minkowski model (Young and Hamer 1987:86):

$$d_L(\mathbf{x}_i,\mathbf{x}_j) = \left(\sum_k |x_{ik} - x_{jk}|^L\right)^{1/L} \tag{2.1}$$

with $L \geq 1$ as a parameter that allows different spatial distances to be used. There are three typical cases when different values of L are considered:

- $L = 2$. This is called Euclidean distance. It is often employed due to its geometrical 'intuitiveness'. When applied in psychology, it often provides good fits to human similarity judgment when objects are holistically perceived (Goldstone 1998). For example, for stimuli with integral dimensions such as hue, saturation, and brightness for color, this would generate a better fit model.

- $L = 1$. This is the Hamming or City-block distance. The distance between two items equals the sum of their dimensional differences. This method is sometimes favored as it is more robust to outlying objects. Compared with the Euclidean distance measure, this model would capture the notion well when stimuli are in separable dimensions (e.g. size and color).

- $L = \infty$. It gives the Chebyshev distance, which reduces the original to the following formula:

$$d_{\infty}(\mathbf{x}_i, \mathbf{x}_j) = \max_k |x_{ik} - x_{jk}|. \quad (2.2)$$

It can be shown that the distance function defined by the Minkowski model (equation 2.1) is a metric (Young and Hamer 1987:91).

Coefficients in Vector Space

Many angular coefficients that measure the proportionality and independence between two vectors in a vector space can also be regarded as a type of similarity measure.

The simplest one is the inner product coefficient:

$$s(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i * \mathbf{x}_j = \sum_k x_{ik} x_{jk}. \quad (2.3)$$

When normalized, it becomes the cosine coefficient:

$$s(\mathbf{x}_i,\mathbf{x}_j)=\frac{\mathbf{x}_i*\mathbf{x}_j}{\|\mathbf{x}_i\|*\|\mathbf{x}_j\|}=\frac{\sum_k x_{ik}x_{jk}}{\left[\sum_k (x_{ik})^2\sum_k (x_{jk})^2\right]^{1/2}} \tag{2.4}$$

A related one is the correlation coefficient:

$$s(\mathbf{x}_i,\mathbf{x}_j)=\frac{\mathbf{x}_i*\mathbf{x}_j}{\|\mathbf{x}_i\|*\|\mathbf{x}_j\|}=\frac{\sum_k (x_{ik}-\bar{x}_i)(x_{jk}-\bar{x}_j)}{\left[\sum_k (x_{ik}-\bar{x}_i)^2\sum_k (x_{jk}-\bar{x}_j)^2\right]^{1/2}} \tag{2.5}$$

The correlation coefficient becomes a cosine coefficient, when the vectors are centered relative to a zero mean, and reduced to have unit variance.

For the purpose of comparison, we list here two classic coefficients that are often employed as similarity measures (Voss and Driscoll 1992, Ruge 1992). They are Jaccard and Dice coefficients, respectively. Later, we will see their corresponding measures for discrete data.

$$s_J(\mathbf{x}_i,\mathbf{x}_j)=\frac{\mathbf{x}_i*\mathbf{x}_j}{\|\mathbf{x}_i\|^2+\|\mathbf{x}_j\|^2-\mathbf{x}_i*\mathbf{x}_j}=\frac{\sum_k x_{ik}x_{ik}}{\sum_k (x_{ik})^2+\sum_k (x_{ik})^2-\sum_k x_{ik}x_{ik}} \tag{2.6}$$

$$s_D(\mathbf{x}_i,\mathbf{x}_j)=\frac{2\times\mathbf{x}_i*\mathbf{x}_j}{\|\mathbf{x}_i\|^2+\|\mathbf{x}_j\|^2}=\frac{2\sum_k x_{ik}x_{jk}}{\sum_k (x_{ik})^2+\sum_k (x_{ik})^2} \tag{2.7}$$

Most similarity coefficients are nonnegative and bounded by unity, $0\le s(\mathbf{x}_i,\mathbf{x}_j)\le 1$, some have the correlational nature satisfying $-1\le s(\mathbf{x}_i,\mathbf{x}_j)\le 1$, and a few, like the simple inner product, are unbounded.

2.3.1.2 Qualitative Data

Binary or logical type

In binary/logical data there are only two values for each attribute dimension, i.e. $x_{ik}\in\{0,\ 1\}$. These coefficients and metrics for the quantitative data still hold. However, the results are greatly simplified and can be presented in different formats (see the discussion below).

Nominal (multilevel) data

Since quantitative models like MDS mainly deal with continuous dimension data, they have in-principle limitations when it comes to nominal variables with several levels – for instance, when a dimension is 'color', the spectral ordering is meaningful for color. But the values of which admit no meaningful serial ordering to humans' perception.

2.3.1.3 Limitations

Aside from inappropriate dealing with nominal data in the continuous dimensions, spatial models are also under attack for the underlying assumption that similarity can be related to the distance in space. Symmetry, one of the four main metric axioms for distance, has been the main focus of the attack by critics. For example, similes such as "butchers are like surgeons" vs. "surgeons are like butchers" would entail different indications towards subjects and targets, which suggest human similarity judgments need not be symmetrical (Medin et al. 1993). Experiments on similarity also demonstrate this effect on directional similarity judgments (Tversky 1977).

2.3.2 Feature-based Models

The feature-based models are designed to overcome some of the difficulties with the spatial models. Instead of representing objects as points in a space with continuous dimensions, the feature-based models represent them as a set of discrete, binary features. They assume that objects are decomposed into underlying features, and compared on the basis of commonality of features. A feature can be any property or attribute of the object. Essentially, the collection of features forms the intent of the object. Similarity is then assessed by measuring the overlap of the feature sets of the compared objects.

In fact, as we will introduce in the next subsection, many classical similarity measures embody the same spirit when the similarity between two objects is concerned. They start from observing the basic characteristic of similarity—number of identical features/attributes shared by two objects. A natural and intuitive prediction is conceived that the similarity increases along with the addition of common features to any two objects.

Using set theory, the attributes sharing information between two objects *a* and *b* can be expressed: $A \cap B$, the attributes that are common to both objects; *A–B*, the attributes that belong to *a* but not *b*; *B–A*, the attributes that belong to *b* but not to *a*. Again, we will discuss various measures for different data types under this set-theoretic notion.

2.3.2.1 Qualitative Data

The feature-based model can describe well many qualitative data similarity analyses.

Binary/Logical Data

Ratio Model

A ratio model can be presented as (Tversky 1977):

$$S(a,b) = \frac{f(A \cap B)}{f(A \cap B) + \alpha f(A - B) + \beta f(B - A)} \tag{2.8}$$

where *f* is a function which is often represented using the cardinality of a set. A specific set-theoretic model can be represented with this generalized model, given a setting for the parameters α and β. We will thus introduce some of the classic similarity measures as the special cases of this ratio model. In all these cases, the function *f* is represented using set cardinality, i.e. $f = | \cdot |$.

1. Jaccard (1908) coefficient, also called Tanimoto (1958) measure. When $\alpha = \beta = 1$, the formula 2.8 can be reduced to:

$$S_J(a,b) = \frac{|A \cap B|}{|A \cup B|} \tag{2.9}$$

This is the simplest form of this class of coefficients.

2. Dice (1942) measure/coefficient. When $\alpha = \beta = 1/2$ the original ratio model can be simplified as:

$$S_D(a,b) = \frac{2 \times |A \cap B|}{|A \cap B| + |A \cup B|} = \frac{2 \times |A \cap B|}{|A| + |B|} \tag{2.10}$$

Compared with the Tanimoto measure, it gives more weight to the matches of attributes than mismatches.

3. Anderberg (1973) measure. This is a less often quoted method. The resultant formula can be obtained from 2.8 when $\alpha = \beta = 2$:

$$S_A(a,b) = \frac{|A \cap B|}{|A \cap B| + 2 \times |A - B| + 2 \times |B - A|} \tag{2.11}$$

This, contrary to the Dice measure, gives more weight to the mismatch of attributes than matched ones.

It is not difficult to verify that the above three formulae have the following relationship:

$$\frac{S_J}{1 - S_J} = \frac{2S_A}{1 - S_A} = \frac{S_D}{2(1 - S_D)} = \frac{|A \cap B|}{|A - B| + |B - A|} \tag{2.12}$$

which indicates that they are monotonic functions of each other. From their parameter settings in the original ratio model, it can be observed that: $S_A \leq S_J \leq S_D$.

There are other similarity measures where the factors of both negative attributes (i.e. $\bar{A}$ and $\bar{B}$) are also considered (Sneath and Sokal 1973:129-135, Gower 1985).

It can be shown that both the Jaccard and Anderberg measures are metrics, while the Dice measure is not (Gower and Legendre 1986).

Contrast Model

Related to the ratio model, Tversky's much-celebrated paper (Tversky 1977) proposed a general contrast model:

$$S(a,b) = f(A \cap B) - \alpha f(A - B) - \beta f(B - A) \tag{2.13}$$

The contrast model expresses similarity between objects as a weighted difference of the measures of their common and distinctive features. This model, together with ratio model, allows for the violations of minimality, symmetry, and triangle inequality axioms on which spatial/geometric models are based. For example, setting parameters α and β differently can model the asymmetry of the directional similarity judgments we discussed before.

Product Model

There is a lesser-known model for binary discrete data called Product Rule (Estes 1994:41-43). Let c and d be the parameters corresponding to common (matched) and distinct (mismatched) attributes of two objects. The product rule is formed by entering the parameter c into the product once for each match and d once for each mismatch:

$$S(a,b) = c^{|A \cap B|} d^{|A-B|+|B-A|} \tag{2.14}$$

This, in fact, can be transformed into a form of the contrast model by taking the logarithm of the expression. The general difference between the two models is that the log of the product similarity yields the linear dependence of the number of common and distinct features. In addition, the product measure is symmetrical whereas the contrast measure is not.

Nominal (multilevel) data

A common strategy to deal with nominal data is to recode them in terms of an expanded set of disjunctive binary codes.

Another simple scheme to treat multilevel qualitative data is to allocate a score s_{ijk} when comparing object i and j on the kth attribute. The simplest rule is to score unity when both objects have the same value on the kth attribute, otherwise score zero:

$$s_{ijk} = \begin{cases} 1 & \text{if } x_{ik} = x_{jk} \\ 0 & \text{otherwise} \end{cases}$$

The overall similarity is then the average of overall scores of each feature:

$$s_{ij} = \frac{1}{m}\sum_{k=1}^{m} s_{ijk} \tag{2.15}$$

Formula 2.15 is essentially an application of the Dice measure in the multilevel data scenario when both objects *i* and *j* have the same set of *m* attributes, and a feature is considered to be a pair (*k*, *x*), meaning that object has attribute *k* with value *x*. Given this interpretation, we have $|A|=|B|= m$ in formula 2.10.

Similarly, the application of Jaccard measure can be described in the form:

$$s_{ij} = \sum_{k=1}^{m} \frac{s_{ijk}}{2 \times m - s_{ijk}} \tag{2.16}$$

When considering the weight w_k of each feature and a more general function of s_k, a very general similarity measure can be formulated (Gower 1971):

$$s_{ij} = \frac{\sum_{k=1}^{m} w_k(x_{ik}, x_{jk}) s_k(x_{ik}, x_{jk})}{\sum_{k=1}^{m} w_k(x_{ik}, x_{jk})} \tag{2.17}$$

This formula can, in fact, allow for different data types for feature similarity determination as well as different determination methods as long as the returned feature similarity is in a standard range of between 0 and 1 (Tudhope and Taylor 1996).

2.3.2.2 Quantitative Data

When the type of a specific feature is a quantitative value, the similarity value can then be measured in distance:

$$s_{ijk} = \frac{|x_{ik} - x_{jk}|}{\max\limits_{p,q}|x_{pk} - x_{qk}|} \tag{2.18}$$

The denominator indicates the range of features *k* can have.

A more general function is given as follows for the overall Minkowski distance:

$$\frac{1}{m}\sum_{k=1}^{m}\frac{|x_{ik} - x_{jk}|^{t}}{r_{k}^{\ t}} \tag{2.19}$$

where r_k is a range function for feature k.

2.3.2.3 Limitations

The original feature-based model (i.e. contrast model) is based on set-theoretic notions. It has the inherited weakness in representing continuous dimensions as well as nominal variables. Although various devices are introduced to deal with this (Tversky 1977), the resultant representations tend to deviate from the original spirit of feature-based representation. For example, by expanding the binary code to multiple values so as to accommodate the nominal variables scenario, some less meaningful features are introduced.

This representational inadequacy, which also occurs with the spatial models, has led critics to develop other similarity models (Hahn and Chater 1997). Moreover, going back to the theory about concept, it is noted that concepts cannot be viewed as mere collections of features. It is the collection of all the features/attributes, plus the interactions between them, that define a concept. Hence, the relationships between the features must be addressed and properly presented. This is particularly evident when we discuss other similarity-related comparison schemes showing that deeper, structural, relational similarity is more fundamental to the overall interpretation and modeling. In human similarity judgment experiments, evidence has been gained that relational properties play an important role in determining similarity (Goldstone 1994, Medin et al 1990, Gentner and Markman 1997). For example, when comparing the similarity between stories, human subjects would rate them more similar for stories that exhibit deeper, structural similarity (e.g. plots) than those that show only attributional similarity (e.g. characters). We use a simple arithmetic analogy as another example. People would consider 3:6 is more similar to 2:4 than to 3:4, although the third has a common feature (the number 3) with the first; the second does not. Again here, the relationship between features plays an important role in the overall similarity determination. Therefore, in modeling similarity, there is a need to take into account *alignment*, a process of creating interdependent correspondences between the features/attributes of compared objects (Goldstone 1996). This is the main starting point for alignment-based models.

2.3.3 Alignment-based Models

From the above discussion of the weaknesses of the spatial and feature-based models, we have seen that they seem incapable of accounting for some of the more complex situations in modeling similarity when the interdependent relationships between the attributes of compared objects need to be addressed. Violations can be observed that are against the basic laws of similarity. One example of this is the existence of nonmonotonicities in similarity ratings. The monotonicity assumption states that when the same feature is added to both objects, their similarity should increase. However, the actual rating may not increase, or even decrease. For example, an added feature of *green* to the color of a car's wheel and a truck's hood may not increase their similarity rating (Goldstone 1998). This is due to a mismatch between the structure and the feature, as the common feature green color is referring to different parts of the objects to be compared. Thus it is highly unlikely to increase similarity by forcing feature comparisons between unlike components. In many occasions, the structure of objects can be better represented hierarchically (objects are embedded in one another) and/or propositionally (objects take other objects as arguments). In all these cases it would be wise not to just simply match features, but to determine beforehand which parts correspond to, or align with, one another (Goldstone, in press).

Inspired by the work on analogical reasoning (Gentner 1983, Holyoak and Thagard 1989), the alignment-based model was developed to accommodate this nonmonotonicity by matching features' influence on similarity more if they belong to parts that are placed in correspondence (Goldstone 1994, 1996, Markman and Gentner 1993). The constraints on correspondence from the work of analogy state that displayed elements tend to be placed in correspondence if they are similar to each other, if they play the same role within their respective objects, and if they are consistent with other correspondences (Goldstone 1996). The alignment-based model includes the notion that correspondences may influence each other. The extra effort is by adding the process of creating interdependent correspondences between the parts of compared objects.

The alignment-based models provide a process account of construction of a similarity comparison. This alignment process gives the model the dynamics necessary to produce nonmonotonicities. From this, we see that the current development of similarity models is more capable of encompassing the complexity of mind, and is moving in the direction of process modeling.

2.3.4 Transformational Models

Another process-oriented model is the transformational model. This approach has a computational origin and places process dynamics at the center. It is based on the assumption that similarity is inversely proportional to the number of operations required to transform one object to the other.

Instead of focusing on the component representation issues, this approach centers on the functional manipulation of objects in a more abstract sense. Thus a representation with this approach will be an account that will encompass virtually any representation. We will introduce two representative approaches within this model.

2.3.4.1 Representational Distortion

One of the main developments of this category is called *representational distortion* (Chater and Hahn 1997). It is based on the intuition that two representations are similar to the extent that there is a simple transformation which distorts one to the other. Similarity is then the estimation of the complexity of this transformation process. The complexity is defined as a computational function in a very general sense by the

Kolmogorov complexity (Li and Vitanyi 1997). The essential idea of Kolmogorov complexity is that the complexity of any mathematical object *x* can be measured by the length of the shortest computer program that is able to generate that object *x*. This length is the Kolmogorov complexity, $K(x)$ of *x*. Thus calculating the representational distortion between two representations *A* and *B*, is determined by the length of the shortest program which 'distorts' *A* into *B*, i.e. the conditional complexity $K(B|A)$. This formulation takes advantage of the rigorous and rich theory of Kolmogorov complexity developed within mathematics and computer science and applies it to the mental representations of psychological similarity judgment.

The model of representational distortion exhibits some attractive and interesting properties (Hahn and Chater 1997). It can be viewed as a generalization of the existing similarity models to the extent that similar sets of features or nearby points in space correspond to short programs. The self-similarity is maximal since no program is required to transform an object to itself. It also allows for asymmetry to be built because of the conditional complexity definition.

This very general representation nature of representational distortion theory can take account of the great flexibility in human similarity judgments. As is often the case, goals and knowledge of the subject may affect the representations which are formed. This flexibility in fact also poses difficulty in modeling the effects of knowledge, which forms as input may radically affect the program length required to transform the objects. Moreover, the approach focuses on a global similarity value and leaves little room for weighting some aspects of local sub-parts representations (Hahn and Chater 1997).

The precise measure of similarity in the notion of representational distortion has yet to be developed to encompass the psychological complexity of humans' minds. The fundamentals of this approach indicate a promising direction for future research.

2.3.4.2 Emergence-based Approach

Another recently proposed model in the spirit of object transformation is the ***emergence-based*** approach to categorization and similarity (Goertzel and Kalish 1996). This view ties the similarity determination with the process of object categorization. "Categorization is emergence-seeking, and similarity is intertransformability" (Goertzel and Kalish 1996). When a collection of objects gives rise to a substantial emergent pattern, then a natural category will be formed by these objects. And similarity is defined if two objects can form a natural category of two, which is equivalent to stating that two objects can be easily transformed one to the other.

In a more general mathematical and conceptual construction called *psynet model of mind* (Goertzel 1997), the theory of pattern provides the basis upon which a new similarity and categorization approach is built. It states that the psychological mind is made of abstract structure, of a pattern. A *pattern* can be characterized as "a representation as something simpler" (Goertzel and Kalish 1996). This notion effectively defines a pattern as a process. Specifically, a pattern is defined as a process p that produces an object $R(p)$ that approximates another object X. A special case of the formalism is when the process p consists of two parts: a 'program' section f and a 'data' section Z. The program p transforms X into Z, and Z into an approximation $R(p)$ of X.

Since similarity depends on finding the emergent pattern as a natural category, one needs first to seek a category within a collection $D=\{D[1], D[2],...,D[N]\}$ of objects. Suppose the output of such an approximation is a subset category $D'=\{D[i(1)],...,D[i(k)]\}$ of the original collection D, we can measure the effect of adding a new object X to the subcollection D'. The amount of 'information added' can be quantified as:

$$J[X|D'] = 1 - \frac{|f(D'+X)| - |f(D')|}{|f(X)|}$$

where | | is a 'simplicity function' mapping the union of the space of objects and the space of processes into the nonnegative real numbers. When $J[X|D']=1$, it means that D'

provides maximal utility to represent X. While $J[X|D']=0$ means that D' is of no use in representing X.

The principle of category formation is then to divide D into mutually exclusive subcollections, where each subcollection is a natural category satisfying that, for X within D', the quantity $J[X|D']$ is large.

In sum, transformational models have abstracted human similarity judgments in a very general sense so that they can succeed in describing and modeling complex scenarios. Since they emphasize process modeling, the crucial point is finding optimal transformational operations.

2.4 Implications

We have presented four major psychological models of similarity, with the inclusions of relevant constructs and formulae that are also applicable in other disciplines. The first two are traditional models that represent the typical *data modeling* approach to similarity, and the last two are in the direction of *process modeling*. Spatial/geometric models are particularly suitable for quantitative data when the dimensions are clear-cut and the values for dimensional data are numerically represented. Feature-based models inherit more intuition from humans' similarity judgments and are capable of dealing with qualitative data. Alignment-based models add a process step to assure that the correspondent parts between objects are really 'comparable'. Transformation models intend to simulate pattern-recognition and transformation operations in humans' cognitive processes. This last model is seen as quite a departure from traditional ones in that it emphasizes the process dynamics of converting one object into the other rather than the static representation of objects only.

There are also some other computational ideas that have been used to model cognitive processes, such as neural network and case-based reasoning. These models are not directly concerned with providing an account of similarity, but similarity is central to the way they operate (Hahn and Chater 1997).

To date, however, there has been no computational system that exhibits the flexibility of accounting for similarity as humans do, the flexibility of choosing the respects/features that are relevant, and the flexibility of weighting them into an overall account. Moreover, other factors, such as context, goals, and even time, can also affect the subjective judgments (Medin and Goldstone 1995). The similarity problem is indeed very difficult to solve, perhaps as difficult as modeling a more fundamental problem – the concept/object itself. Researchers have also argued them as the chicken and egg problem (Hahn and Chater 1997).

This thesis is not aimed at developing another general similarity model. Rather, it attempts to address a specific problem (semantic similarity) in a less general domain. Therefore, the ideas and principles from the above discussed general models can be absorbed and applied. For example, we see the representation in feature-based models as a proper approximation to a human's intuition process. The arrangement of features/respects in alignment-based models can ensure an effective matching. The transformational approach provides an account of the guidance of operational processes.

Chapter 3

Defining Similarity

3.1 Introduction

From various domains and complex similarity models, we go back to the single and fundamental issue about similarity, its definition. From mathematics to psychology, from ecology to artificial intelligence, many similarity formulae have been contrived to define similarity under various background knowledge and domains. However, many of these are restricted to some specific domain, and were proposed without much explicit theoretical justification. Most often, similarity formulae are directly presented as definitions, rather than derived theorems based on certain assumptions. We therefore seek, under a most generic scenario, a quantitative definition that is both theoretically sound and widely applicable. This would require a high degree of abstraction of the domain and some intuitively appealing assumptions. In particular, rather than directly stipulating a similarity formula, we use set-theoretic notions to derive a similarity definition that is based on certain well-accepted assumptions. The derived results can be seen as a generalization of some of the classical definitions of similarity.

In order to derive proper similarity measures, we present a framework for the conceptualization of universal object definitions and comparisons that is based on the classical cognitive and psychological views about concepts and categories. We then argue that this framework provides a general account for modeling object comparisons, and some of the specific comparison schemes can be further abstracted and quantified using information-theoretic notions so that a simple computational means of measuring universal object similarity can be achieved.

In what follows, first a framework for the conceptualization of a universal object definition and comparison scheme is presented. Then, some general similarity definitions are derived based on this framework and additional well-accepted assumptions with regard to humans' similarity judgment processes. This is followed by a discussion and clarification of other similarity related concepts and the relationships among them. In section 3.3, the proposed general definition is modeled in information-theoretic terms to arrive at some operational similarity measures for the domains where a probabilistic model holds. These specific measures then become the major building blocks for theory development in the remaining parts of this thesis. In the last section, we argue that one of the defined similarity related concepts can be treated as an approximation to the transformation-based models discussed in the last chapter.

3.2 Object Comparison and Measurement

We start with a feature-based approach to modeling objects, but we wish to avoid dependence on how features are chosen, how many features are present, and how to align features. Therefore, we identify with each object a single feature, which we will denote its *content*. The content of an object encapsulates all the attributes of the object as a whole, and it uniquely identifies the object. In the following sections, we develop a computational means of determining universal object relationships based on this model by adapting many of the measures described in Chapter 2 for feature-based and other models.

3.2.1 Object Representation and Measurement

Before we discuss the comparison between objects we need to understand well the definition of an object. This includes two accounts: the representation of an object and measuring this representation in a quantitative manner.

For object representation, we follow the classical view described in section 2.2.1 that objects/concepts have well-defined boundaries and are describable by sets of necessary and sufficient conditions (i.e. dual properties of intent and extent). This classical model is both elegant in expression and rigorous in logic. Particularly, given the mathematical properties from set theory, there are certain operations by which the content representation can be measured and manipulated.

This classical view has been quite successful in modeling relatively simple and concrete objects in the physical world. However, it has its weakness in modeling complex objects, especially those that designate nonphysical domains — emotions, language, social institutions — that describe abstract relations in humans' mental cognitive and reasoning processes. Since the 1950s, the adequacy of this classical model has been questioned (Wittgenstein 1953, Rosch and Mervis 1975), and various alternative theoretical views have proposed, among them the *prototype view*, the *exemplar view*, the *frame view*, and the '*theory view*'. For example, prototype views emphasize the gradedness, prototypicality, or goodness-of-example in objects' extensional phenomena. That is, one of its extensional members may have a higher degree of 'representativeness' of the whole concept. An example is the concept of *mother*. There need be no necessary and sufficient conditions for motherhood shared by normal biological mothers, donor mothers, adoptive mothers, foster mothers, and so on. However, by social stereotypes, the housewife-mother concept defines cultural expectations about what a mother is supposed to be. Such goodness-of-example yields prototype effects (Lakoff 1987:74-80). The prototype view also questions the strict boundaries of concept and instead concerns the object's extensional vagueness. Lakoff expanded the prototype view by highlighting human interaction and imagination in the concept formation process—the gestalt perception, metaphoric mapping, metonymic prototyping, radial structure, etc. (Lakoff 1987).

In spite of the apparent advantages of various alternative models, the classical model is still dominant in practical operational use. One explanation for this is perhaps the alternative models are very difficult to interpret in computational terms[3]. This, however, is crucial in realizing object measurements. Together with the consideration of keeping our rule of maximal abstraction, we consider the classical set-theoretic view as an appropriate candidate in object representation, especially in fulfilling our task of modeling lowest level of simple and lexicalized concept representation.

In the fashion of maximal abstraction, we go beyond the component point of view of object representation. We do not restrict the types of characteristics that define an object as long as they are used consistently to describe all the objects in the object space. For instance, the object representation can be characterized as a whole by its relations with other objects or its position in the whole object space, rather than by an aggregation of its intensions or extensions. For example, in a process sense, an object can be represented by its complexity using the theory of algorithmic complexity (Li and Vitanyi 1997).

Once we have a specific form representing an object, we can then design a scheme that quantifies this representation. This is in the belief that in the universe of objects, each object contains a certain amount of *content* or *information* that characterizes itself as a single entity, and at the same time, differentiates itself from other objects. For the set-theoretic representation, the content of an object can be measured by a function of its subcomponents, attributes, features, respects, etc. For the algorithmic complexity model, the content is the value of complexity that corresponds to the length of the shortest computer program that is able to generate (the description of) that object.

Under the most general situation we will use the capital letter A to refer to the representation of an object a. Its measurement, the amount of content or information, is defined as $I(A)$. Obviously, the content measure is an absolute measure whereas the unit of it depends on a specific representation.

[3] One exception may be the frame view, where it has been integrated into knowledge representation in artificial intelligence study. However, this is at the high cost of manually building knowledge base and the applications tend to be limited to certain domains.

Generally, the object measure function I is a nonnegative interval scale function. There are certain assumptions about this function:

- Nullity. $I(A) = 0$ iff $A = \phi$.
- Monotonicity. $I(A) \geq I(B)$ if $A \supseteq B$.
- Feature Additivity. $I(A \cup B) = I(A) + I(B)$, if $A \cap B = \phi$.

These are straightforward assumptions. They can be verified by the simple function of set cardinality.

We present a simple example to illustrate a potential object representation and its measurement. The example is a simple geometry picture, and similar pictures will be used throughout this chapter to illustrate other concepts. Using set-theoretic notion, we can represent this object by its subcomponents with three kinds of features or attributes for each of the three elements: shape, shading, and location. Assuming equal weight for each feature and set cardinality as the measurement function, the content value for the object is calculated as 9.

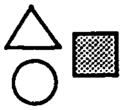


Figure 3.1 A geometric picture object

3.2.2 Object Comparison

Once there is a certain way to define a representation of an object and its measurement, we can then explore the relationship between two objects when they are considered together. Typically we would like to see how close or how different two objects are in an abstract scheme. Before we introduce the appropriate similarity measure we first discuss some relevant concepts that are essential to object comparison. Since we use set-theoretic notions to represent an object, the concepts we introduce here will be directly related to those concepts in typical set operations.

Unification

The first measure in describing the relationship between two objects is *unification.* It refers to the content or information for the whole combined part that two objects possess. In set-theoretic terms, it is the content measure of the union of two objects. The definition is as follows:

$$unification(a,b) = I(union(a,b)) = I(A \cup B) \qquad (3.1)$$

where $union(a,b)$ is a proposition that represents the joint occurrence of a and b.

Commonality

Corresponding to set intersection, the content measure for an intersection of the representations of two object is called *commonality.* In other words, the commonality of two objects refers to how much content or information they share in common in their representations. This is defined as follows:

$$commonality(a,b) = I(common(a,b)) \qquad (3.2)$$

where $common(a,b)$ is a proposition that states the commonalities between a and b. In set-theoretic terms, the above can be further represented as:

$$commonality(a,b) = I(common(a,b)) = I(A \cap B) \qquad (3.3)$$

For example, if a is the object *table* and b is *chair*, the parts that they share in common are a piece of furniture, and leg(s) to support the top or seat. A simple case for the value of commonality could be just the sum of the number of identified common features.

Since we are discussing the relationship between two objects only, we can abbreviate the above expression as **comm** or **C** when the objects considered are clear from context.

It can be observed that the range of commonality is between 0, when there exists no common parts between two objects, i.e. $A \cap B = \phi$; and maximum when one actually contains the other, i.e.

$$\max(commonality(a,b)) = \min(I(A), I(B)) \tag{3.4}$$

Therefore,

$$0 \leq comm \leq \min(I(A), I(B))$$

Notice that commonality is an absolute measure in the comparison of two objects.

Difference

The complement concept for commonality is *difference*. Therefore the difference of two objects is defined as the amount of content resulting by subtracting the commonality of two objects from their unification:

$$\begin{aligned} difference(a,b) &= unification(a,b) - commonality(a,b) \\ &= I(A \cup B) - I(A \cap B) \\ &= I((A \cap \overline{B}) \cup (A \cap B) \cup (\overline{A} \cap B)) - I(A \cap B) \\ &= I(A \cap \overline{B}) + I(A \cap B) + I(\overline{A} \cap B) - I(A \cap B) \\ &= I(A - B) + I(B - A) \end{aligned} \tag{3.5}$$

The third equation is based on the feature additivity assumption, since $A \cap \overline{B}$, $A \cap B$ and $\overline{B} \cap A$ are disjoint. Difference is thus the amount of content from each object that is not shared.

Notice that difference as defined here corresponds to the set-theoretic definition of 'symmetric difference.' We choose to adopt this definition so that we can treat difference as a metric distance (see Section 3.4).

Using the same *table* and *chair* example, we can determine the difference between them. For instance, a chair normally has a back. The seat of a chair is for people to sit, while the top of a table usually supports articles. Again, a simple addition of the number of these distinct features would compose a difference calculation.

Similarly to the notion of commonality, we can simplify the notion of the difference between two objects as **diff** or **D**.

There is a relationship between commonality and difference, presented as follows:

$$difference(a,b) + 2 \times commonality(a,b) = I(A) + I(B) \tag{3.6}$$

The proof is trivial and thus is omitted.

Some observations about the range of difference can be easily obtained. The difference reaches its maximum when commonality is 0, i.e.

$$\max(\text{diff}) = I(A) + I(B) \tag{3.7}$$

and its minimum when commonality is maximal, i.e.

$$\min(\text{diff}) = I(A) + I(B) - 2 \times \max(\text{comm}) = |I(A) - I(B)| \tag{3.8}$$

Therefore,

$$|I(A) - I(B)| \leq \text{diff} \leq I(A) + I(B)$$

Like commonality, difference is also an absolute measure of relatedness of two objects.

Another view of formula (3.6) is that:

$$\text{diff} + \text{comm} = I(A) + I(B) - \text{comm} \tag{3.9}$$

Similarity

Given the above two definitions regarding the absolute sense of object overlapping relationships, we can now define *similarity* — a relative measure of closeness or relatedness of two objects. It essentially corresponds to the *commonality* in the absolute measures, but in a 'relative' sense. Since it is a relative association measure, it should be independent of the unit used in the content determination.

In the following we present some basic assumptions for a valid similarity measure. They follow common descriptions about the characteristics of a similarity measure (Gower 1985) and cover the essential schemes we described in section 2.2.2 for the conventional assumptions about similarity. We will then derive the similarity definitions from these assumptions. We use the abbreviations C, D and S to denote commonality, difference, and similarity, respectively. And R^+ refers to a non-negative real number.

(A0) $S(C,D)$: $R^+ \times R^+ - \{(0,0)\} \to R^+$, for $C, D \in R^+$

(A1) $0 \le S(C,D) \le 1$, for all C and D

(A2) $S(C, 0)=1$, for $C>0$

(A3) $S(0, D)=0$, for $D>0$

(A4) $S(C_1,D_1) \ge S(C_2,D_2)$, for $C_1 \ge C_2$ and $D_1 \le D_2$. More strongly, $S(C_1,D_1) = S(C_2,D_2) + k_1(C_1 - C_2) - k_2(D_1 - D_2)$, for $k_1, k_2 \in R^+$.

Assumption 0 essentially defines similarity (S) as a function of commonality (C) and difference (D) of two objects that yields a non-negative real number. Obviously, the non-event scenario $\{0,0\}$ should be ruled out.

Assumption 1 further requires that the values of the similarity measure be in the range of $[0,1]$. Most similarity measures or coefficients satisfy this constraint.

Assumption 2 indicates that the similarity value reaches its maximum of 1 when there is no difference between two objects, i.e. D=0. In this case, the two objects are identical. Commonality is then equal to either one of the objects' content: $C = I(A) = I(B)$.

Conversely, Assumption 3 states that the similarity value reaches its minimum when there is nothing in common between two objects, i.e. $A \cap B = \phi$. In particular, this minimum value is set as equal to 0.

Assumption 4 reflects that similarity is a monotonic function of commonality and difference, i.e. the more in common two objects share and the less they have in their distinct parts, the more similar they are. The further stronger condition restricts this monotonic relation to a linear one, i.e. similarity is linearly increasing as a function of both increased commonality and decreased difference. In manipulating objects during the comparison, two specific but separate schemes can be followed:

(A4a) $I(A) + I(B) =$ constant, i.e. the sum of the amount of content each object contains is unchanged during the comparison process;

(A4b) $I(A \cup B)$ = constant, i.e. the unification is unchanged.

These two schemes define approaches for normalization when comparing similarity among diverse pairs of objects — by keeping $I(A) + I(B)$ or $I(A \cup B)$ constant, difference (D) becomes a negative function of commonality (C) (see equations 3.6 and 3.5) as indicated in the assumption.

The scenario corresponding to Assumption 4a can be visualized in Figure 3.2. We just change the overlapping scheme between the two objects, while keeping the total content of each unchanged. In Figure 3.2, the content of object b is repositioned by moving it apart from A, thus the overlapping part between A and B is smaller.

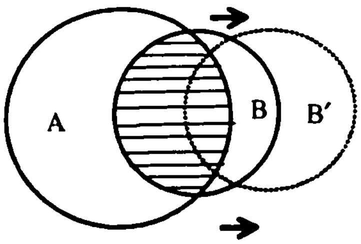


Figure 3.2 Illustration of Assumption 4a by repositioning the content of object *b*

The geometric pattern example to illustrate this type of manipulation is given in Figure 3.3. Here, one of the common features between *a* and *b*, the shading of the circle, is transformed from a non-shaded to a shaded one in object *b′*, while the overall content value of each is unchanged. Again, if we use set cardinality to calculate the content, we have $I(A)+I(B)=18$. The commonality reduces from 7 to 6, and the difference increases from 4 to 6.

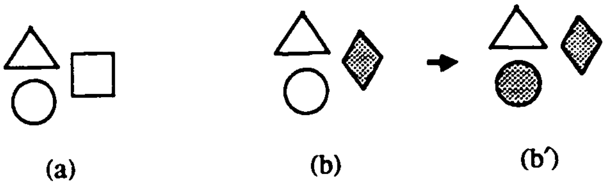


Figure 3.3 Example of objects manipulation under the condition of Assumption 4a

The scenario corresponding to Assumption 4b can be illustrated in Figure 3.4. In this figure, only the curve of *B* along the intersection part is 'stretched' left so that the commonality becomes larger (and hence the difference becomes smaller) while the overall area of *A* and *B* together, i.e. $A \cup B$, remains the same.

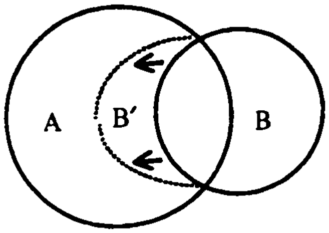


Figure 3.4 Illustration of Assumption 4b by expanding the content of object *b*

A similar geometric pattern example to illustrate this type of manipulation is given in Figure 3.5. In this example, more content features from object *a*, a blank circle, are added to object *b*, while the overall content value that two objects possess together is unchanged. In set cardinality terms, $I(A \cup B) = 11$. The commonality increases from 4 to 7, while the difference decreases from 7 to 4.

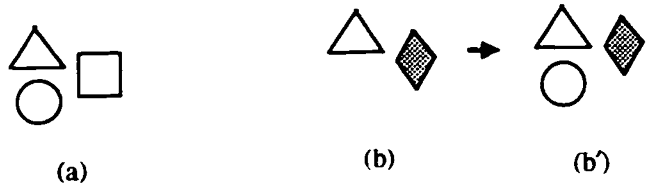


Figure 3.5 Example of objects manipulation under the condition of Assumption 4b

For each of the two object manipulation schemes in Assumption 4, we can derive a similarity definition. Again, for simplicity, we use **sim** or **S** to refer to two objects' similarity: *similarity*(*a*, *b*).

Similarity Theorem 1:

A definition of similarity S that satisfies A0-A4 and A4a is:

$$S = \frac{2 \times C}{I(A) + I(B)} \tag{3.10}$$

Proof: The stronger assumption (A4) indicates that similarity is a linearly increasing function of both increased similarity and decreased difference:

$$S(C_1, D_1) = S(C_2, D_2) + k_1(C_1 - C_2) - k_2(D_1 - D_2), \tag{3.11}$$

where $k_1, k_2 \in R^+$. Given the proposition (A4a) and formula (3.6), we have:

$$D + 2 \times C = I(A) + I(B) = \text{constant}. \tag{3.12}$$

Therefore, for the two states in the similarity manipulation, we have:

$$D_1 + 2 \times C_1 = D_2 + 2 \times C_2,$$

i.e.

$$D_1 - D_2 = -2 \times (C_1 - C_2)\,.$$

Substituting the above into formula (3.11), we have:

$$\begin{aligned} S(C_1, D_1) - S(C_2, D_2) &= k_1(C_1 - C_2) + k_2(2 \times (C_1 - C_2)) \\ &= (k_1 + 2 \times k_2)(C_1 - C_2) \end{aligned}.$$

This is equivalent to saying that there is a linear relationship between similarity and commonality. Without loss of generality, we can express it as:

$$S = k \times C + m, \tag{3.13}$$

where k and m are real value parameters to be determined.

Consider the scenario of (A3), i.e. $S(0,D)=0$ and $C=0$. Substitute these conditions into formula (3.13), we can then obtain $m=0$.

Similarly, with (A2) we can have $S(C,0)=1$ and $D=0$. Substitute the latter into equation (3.12), we obtain $C=\text{constant}/2$. Together with $m=0$, we can obtain k from formula (3.13) as

$$k = \frac{1}{\text{constant}/2} = \frac{2}{I(A) + I(B)}\,.$$

With the resolved parameter values in equation (3.13), we obtain formula (3.10).

Similarity Theorem 2:

A definition of similarity S that satisfies A0-A4 and A4b is:

$$S = \frac{C}{I(A) + I(B) - C} \tag{3.14}$$

Proof: The proof is very similar to the proof for Theorem 1. With the stronger assumption (A4), we have:

$$S(C_1, D_1) = S(C_2, D_2) + k_1(C_1 - C_2) - k_2(D_1 - D_2)\,, \tag{3.15}$$

where $k_1, k_2 \in R^+$. Given the proposition (A4b) and formula (3.5), we obtain:

$$D + C = I(A \cup B) = \text{constant}. \qquad (3.16)$$

Therefore, for the two states in the similarity manipulation, we have:

$$D_1 + C_1 = D_2 + C_2,$$

i.e.

$$D_1 - D_2 = -(C_1 - C_2).$$

Substituting the above into formula (3.15), we have:

$$\begin{aligned} S(C_1, D_1) - S(C_2, D_2) &= k_1(C_1 - C_2) + k_2(C_1 - C_2) \\ &= (k_1 + k_2)(C_1 - C_2) \end{aligned}.$$

Again, this is equivalent to saying that there is a linear relationship between similarity and commonality. Without loss of generality, we can express it as:

$$S = k \times C + m. \qquad (3.17)$$

Similar to the proof in theorem 1, the value for parameter m is obtained as m=0 when (A3) is considered.

With (A2), i.e. $S(C,0)$=1 and D=0, we then have the maximal possible value for similarity in equation (3.16), C= constant. Together with m=0 we can obtain k from formulas (3.17) and (3.9) as

$$k = \frac{1}{\text{constant}} = \frac{1}{I(A) + I(B) - C}$$

With the resolved parameter values in equation (3.17) , we obtain formula (3.14).

When object content measure I is replaced with a set cardinality calculation, the two derived similarity definitions in fact correspond to the Dice measure and the Jaccard measure, respectively (see section 2.3.2.1). Our formulae are more general in the sense that we allow for a general function for object content determination.

Discrimination/Dissimilarity

The complement of similarity between objects is *discrimination* or *dissimilarity*. Since similarity has a limited range, it would be wise to define discrimination in the same manner:

$$\text{discrimination}(a,b) = 1 - \text{similarity}(a,b)$$

Distance

Another widely used relative association measure is *distance*. It is similar to the dissimilarity measure but with a wider range of values. Distance can be seen as the reverse scenario of similarity by projecting the [0,1] range of similarity values to a non-negative real-valued number [0, ∞). Thus we can define distance as:

$$\text{distance}(a,b) = \frac{1 - \text{similarity}(a,b)}{\text{similarity}(a,b)} = \frac{\text{discrimination}(a,b)}{\text{similarity}(a,b)} \qquad (3.18)$$

It is essentially the same as the conventional definition of similarity based on distance:

$$\text{similarity}(a,b) = \frac{1}{1 + \text{distance}(a,b)} \qquad (3.19)$$

It can be seen that the distance shrinks to 0 when similarity is 1. It goes to infinity when there is no similarity.

In a similar fashion, there are other types of distance functions that satisfy the mapping scheme (Batagelj 1995). For example:

$$\text{distance}(a,b) = -\log(\text{similarity}(a,b))$$

Notice the distance measure discussed here is different from the metric sense of distance defined in section 2.2.2, although many distance measures do satisfy the metric requirements.

3.2.3 Relations Among the Comparison Measures

Based on the above definitions and derivations of object association measures, we can uncover and derive some new relations among them.

Figure 3.6 depicts a three-dimensional graphical relationship between similarity and the two absolute content measures, commonality and difference (assuming $I(A) > I(B)$), under the condition that both values of $I(A)$ and $I(B)$ are unchanged.

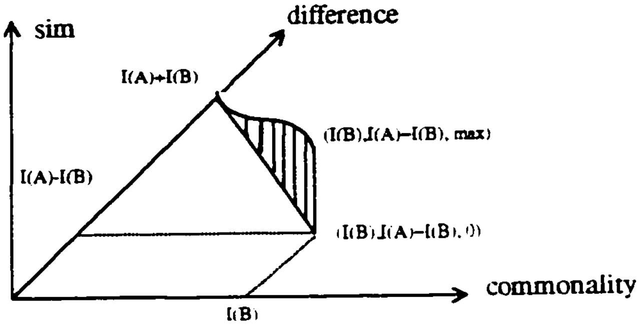


Figure 3.6 Potential range of similarity values regarding the variables *commonality* and *difference*

The striped plane indicates the region of values that similarity can have with respect to the changes of commonality and difference. The plane itself conforms to the constraint of equation (3.6). The curve connecting points $(0, I(A)+I(B), 0)$ and $(I(B), I(A)-I(B), \text{max})$ in that plane would satisfy the general monotonic relationship between similarity and commonality.

The relationship between distance and the two absolute measures are as follows:

$$\text{distance}(a,b) = k \times \frac{\text{difference}(a,b)}{\text{common}(a,b)} \tag{3.20}$$

where k is a positive number denoting the scale of contribution for each absolute measure towards the overall distance value. When k=1, it is the Jaccard measure, and k=1/2 is the Dice measure.

Table 3.1 summarizes the features of the four major measures and their relationships.

	Sameness	Contrast
Absolute measure	commonality	difference
Relative measure	similarity	distance

Table 3.1 Characteristics of four major object comparison schemes

3.3 Information-theoretic Approach to Object Comparison

Feature matching based on set-theoretic notions requires that we extract and interpret a limited list of relevant features from object descriptions where comparisons are conducted. This process is normally task-specific and requires extensive knowledge of the domain. It becomes much harder in trying to replicate this process computationally, considering many factors that may affect the comparison process when humans perform the same task. For example, in determining the saliency for a feature there is a diagnostic factor that refers to the classificatory significance of features (Tversky 1977). We use some examples to explain this. Let's compare the objects *chair* and *stool*. One feature of both is the seat/top where people can sit. At this point, we do not normally consider the number of people that would sit on a chair or stool. When the object *bench* is added to the comparison, that factor would become salient. Similarly, in our simple geometry picture examples, the 'location' feature can be ignored as long as there are always three elements positioning the same spot in a picture.

Another apparent limitation of the feature-based models is the expressive power of capturing all the relationships between features rather than a simple linear combination of feature lists. This is typically required in describing complex objects. More sophisticated models such as prototype, frame-based methods are able to accomplish this in a fashion close to humans' ability and flexibility in describing knowledge. However, developing principled processing accounts for these complex models and structures constitutes a formidable challenge (Barsalou and Hale 1993, Hampton 1993), as this would entail huge amount of work in knowledge extraction, verification and representation. There is also a need of significant computational efforts in terms of the cost of storage, classification, and learning knowledge. Compared with feature-based models, the development of representational formalisms is less tractable in these sophisticated models.

In general, component-based models, whether classical or the more advanced approaches, require explicit elicitation and representation of all *relevant* components of an object before a comparison takes place. By 'relevant' we mean it is situation, content and goal dependent. As we have argued, however, both the availability of this structured knowledge representation and the capacity of processing this to the extent that will reflect actual humans' processes remain to be resolved in order to generate a computational means for universal object comparisons.

Along with the principle of maximal abstraction, we seek an alternative to this by directly generating a global object content measurement on the basis of object relationships in the complete object space. In many domains of application where a probabilistic model holds, the content of an object can be represented by how much *information* it contains. The amount of information an object carries can be measured in a probabilistic way given that it forms a certain probabilistic distribution in the object space. According to information theory, the amount of information conveyed in an object is in reverse to its frequency, i.e. the more common an object is, the less information it would contain. The more specific an object is, the more information it contains.

One way to characterize this is to embody the *Information Content* (IC) concept (Ross 1976). Following the notation in information theory, the information content (IC) of an object can be quantified as follows:

$$IC(A) = -\log P(a)\,, \tag{3.21}$$

where $P(a)$ is the probability of encountering an instance of object a.

Given the information-theoretic approach to defining an object, we can then further derive the measurement schemes in object comparison using this approach.

The *commonality* of two objects a and b would be:

$$commonality(a,b) = -\log P(common(a,b)) \tag{3.22}$$

Similarly, through definition (3.5), the difference of two objects can be expressed as:

$$difference(a,b) = -(\log P(a) + \log P(b) - 2 \times \log P(common(a,b))) \tag{3.23}$$

As for the similarity measure, given equation (3.10), it follows that:

$$Dice\_sim(a,b) = \frac{2 \times \log P(common(a,b))}{\log P(a) + \log P(b)} \tag{3.24}$$

Similarly, the information theoretic measure for similarity with respect to equation (3.14) is:

$$Jaccard\_sim(a,b) = \frac{\log P(common(a,b))}{\log P(a) + \log P(b) - \log P(common(a,b))} \tag{3.25}$$

We therefore provide a simple operational means of quantifying these object comparison schemes. Once we have the correct interpretation of commonality for a specific domain, we can then obtain the value for this information-theoretic similarity measure. The models in Chapter 4, when the problem of lexical semantic similarity is considered, provide a solution to determining the commonality of two objects.

In fact, given the information-theoretic notion of measuring an object's content, we also consider the two absolute comparison measures (commonality and difference) as a way of interpreting object similarity/distance. The next section will detail this view.

3.4 'Difference' as Information Distance

This section describes an informal model of viewing the 'difference' measure (formula 3.5) in object comparison as a measure of absolute information distance between them. The purpose of this is to seek an approximation of the similarity determination in the principle of transformation-based models.

In section 3.2.1, we mentioned that the content of a single object *a* can be measured by the notion of its algorithmic/Kolmogorov complexity: *K*(*a*), which is the length of shortest computer program that is able to generate object *a*. It is then desirable to have a similar measure of absolute information distance/similarity between individual objects. Intuitively, this minimal information distance between object *a* and *b* would be the length of the shortest program to transform *a* into *b* or *b* into *a*. Suppose, for example, that *a* represents category *chair*, and *b* represents category *bench*, then the length of such a program would be significantly shorter than that when *b* represents category *apple*. This is because many of the aspects of the former two objects will be shared, therefore there is no need in the program steps to perform the extra transformations for the common parts. Essentially, what the program does is transform those content representations (e.g. features) that are distinct between them. This is very similar to the notion of the 'difference' measure in our formalization. Transformations are thus the process of deletions and additions of object content one by one among the parts that are different.

This information distance measure is shown to be, up to a logarithmic additive term, equal to the maximum of the conditional Kolmogorov complexities $K(a|b)$ and $K(b|a)$ (Li and Vitanyi 1997:537-542). This formula is asymmetric and thus can explain some of the asymmetric features in object comparison. It can also be modeled to represent symmetry by the average of the distances in either direction, $(K(a|b)+K(b|a))/2$.

Thus it can be observed that the 'difference' measure can be treated as a similarity measure from the view of transformation-based models. As for a specific implementation, say the probabilistic information content determination discussed in the last section, 3.3, we are not clear whether the 'difference' measure calculated from an information content value would qualify as a true representation of a similarity measure. This is because there is no proof that the information content value defined in that way is equivalent to a minimal quantity of information (i.e. shortest computer program) required to represent an object. Nevertheless, this warrants further exploration and perhaps empirical verification. In many parts of the thesis, we are in fact seeking empirical support for this argument. From the literature, similar views are observed by modeling 'difference' as a way of calculating similarity. For example, Hirst (1995) proposed a model of adducing differences and contrasts in objects from lexical taxonomy as a way to determine similarity between near-synonymous words.

3.5 Summary

In this chapter, we constructed a new formulation of general similarity theory. To lay the foundation for similarity determination, we first presented a framework for modeling object comparison using set-theoretic notions. Some of the very fundamental concepts such as object representation, measurement, unification, commonality, difference, are defined and clarified. In particular, we separate object representation and its measurement, which gives rise to greater flexibility in modeling either concept. Together with a set of certain well-accepted assumptions, this framework has led to a derivation of some general similarity definitions. The derived results can be seen as a generalization to some of the classical definitions of similarity. We also presented various general similarity-related concepts, and their relationships are explored.

These object comparison schemes, which are based on the component approaches to object representation, provide a general account for modeling object relationships. However, a concrete operational vehicle often lacks due to the difficulty in fully extracting and representing the knowledge that defines its subcomponents and their relationships among them. To facilitate a feasible computational effort in determining the universal concept similarities, we demonstrated how these established object comparison concepts can be quantified in information-theoretic terms. In later chapters, we will see the effectiveness of these concepts in modeling lexical semantic similarity. In the last section, we argued that the 'difference' concept can be treated as an approximation to the transformation-based models discussed in the preceding chapter.

Chapter 4

Lexical Semantic Similarity

4.1 Introduction

In this chapter we will see how the information-theoretic modeling of object content and the similarity comparisons derived in the last chapter are applied to the central topic of this thesis: lexical semantic similarity — a measure of semantic/conceptual similarity between pairs of lexicalized concepts represented in words or terms. As is often the case for many tasks in psycholinguistics and NLP, a job is decomposed to the requirement of resolving the semantic relation between lower level constituents such as words or concepts. One needs to come up with a consistent, widely applicable computational model to assess this type of relation.

Instead of directly applying the derivations developed earlier to the lexical semantic similarity problem, we adopt an approach that targets it from a different view. This is the area where conceptual similarity is sought in a semantic network. We first discuss the threads of development in the literature in the area of general semantic similarity determination from purely probabilistic approaches with the help of corpus data, to the use of manually-built knowledge sources such as taxonomies or thesauri. Then we develop a new enhanced measure which is an integration of previous models that use information from a hierarchical structure (e.g. taxonomy). We will further reveal how the proposed method aptly fits into the theory of similarity we developed in chapter 3. Essentially, the taxonomic structure provides a good solution to the problem of determining 'commonality' between two arbitrary objects (section 3.3).

The proposed measuring scheme takes advantage of both the corpus-based statistical approach and the knowledge-based taxonomy approach. The taxonomy/lexicon knowledge provides a qualitative and generative framework in measuring concept similarity, while corpus statistics enhances and fine-tunes this qualitative measurement by providing statistical evidence. The proposed model has a practical appeal as it is easy to implement, it does not require a complex and sophisticated knowledge base, and it is domain independent.

The organization of this chapter is as follows. In the next section, we first discuss the concept of semantic similarity and its more general concept — semantic association. In section 4.3, we briefly revicw two main approaches to semantic similarity based on the usage of knowledge as the source of information. After introducing the recently developed and broad-coverage lexical taxonomy WordNet, in section 4.5, we describe the major development in modeling semantic similarity that take advantage of the semantic knowledge which can be extracted from a taxonomy like WordNet. Following that we present a new similarity measure, which can be viewed as an integration of previous methods. In section 4.7, we provide a unified view of how various lexical semantic similarity models (including the proposed one) fit into the general similarity theory that we developed in the last chapter. In section 4.8, experiments are conducted to evaluate various computational similarity models by comparing them against human subject judgment on the ranking of similarity between word pairs. Finally, we discuss related work and summarize the results.

4.2 Semantic Associations

When a word-level semantic relation/association requires exploration, there are many potential types of relations that can be considered: *hierarchical* (e.g. IS-A or hypernym-hyponym, part-whole, etc.), *associative* (e.g. cause-effect), *equivalence* (synonymy), etc. Table 4.1 lists a classification of common types of semantic relations[4]. Among these, the

[4] For a broader view of lexical relations (including morphological and syntactic relations) , readers are referred to Nutter (1989) where over 100 lexical relations are reported in a rich hierarchical structure.

hierarchical relation represents the major and most important type, and has been widely studied and applied as it maps well to the human cognitive view of classification (i.e. taxonomy)[5]. The *IS-A* relation, in particular, is a typical representative of the hierarchical relation. It has been employed to study a special case of semantic relations — semantic similarity or semantic distance (Rada et al. 1989). In this study of semantic similarity, we will take this view, although it excludes some potentially useful information that could be derived from other relations. Strictly speaking, the IS-A relation corresponds to the *class-subclass* and *genus-species* relations listed in Table 4.1. In many occasions, however, the *class-member* relation is also considered as a type of IS-A relation. We therefore do not differentiate them in this study.

General Category	Relation	Example
Hierarchical Relations	Genus - Species Class - Subclass Class - Member(Instance) Whole - Part Object - Attribute	vertebrate - mammal substance - fluid capital - Ottawa bus - seat rose - fragrance
Associative Relations	Co-ordinate Situational Diagonal Genetic Cause - Effect Instrumental Material Concurrent use Paradigmatic Quasi-synonym	social equality - economic equality physician - hospital scheduled castes - bonded labors judge - judgment dispute - litigation slander - speech punishment - prison justice - law moon - lunar genetics - heredity
Equivalence Relations	Synonyms Antonyms Abbreviations Translation Spelling Differences	valve - faucet fertile - sterile compact disk - CD anticipate - aspettarsi behaviour - behavior

Table 4.1 A classification of semantic relations

[5] In modeling human's cognitive view of categorization/classification, Rosch (1978) and Lakoff (1987) have developed alternative approaches like radial, metaphoric and metonymic models in their prototype theory. Also in studying taxonomic hierarchy, they discovered that certain middle levels of the hierarchy (i.e. 'basic-level') have more impact in human knowledge acquisition and organization. For example, in the furniture-chair-rocker hierarchy, chair is the most recognized and well-mastered concept. Unfortunately, many of these models lack a computational means to interpret them.

4.3 Approaches to Lexical Semantic Similarity

The study of words/terms relationships can be viewed in terms of the information sources used. The least information used are *knowledge-free* approaches that rely exclusively on the corpus data themselves. Another approach is so called the *knowledge-weak* approach that utilizes existing pseudo-knowledge bases, such as machine-readable dictionaries (MRD), lexicons, thesauri, and taxonomies whereby inherited rich semantic information can be explored.

4.3.1 Knowledge-free (Corpus-based) Approach

Corpus-based models are typically knowledge-free, which means one forgoes attaining any true 'understanding' of the text and instead concentrates on learning the low-level information that has previously been created by hand. Statistical principles are mainly employed in this approach as the guidelines to the extraction of lexical/semantic knowledge from the corpora.

One of the earliest corpus-based studies of term relationships is by means of the concept of mutual information (MI). MI is an information-theoretic measure of gauging the 'relatedness' between two words (Fano 1961). The definition of MI is as follows:

$$MI(x,y) = \log\frac{P(x,y)}{P(x)P(y)}$$

Intuitively, the probability of seeing *x*, *y* together, *P*(*x*,*y*), gives some idea as to how related they might be. This is adjusted by taking into account the individual probability to correct the problem when *x*, *y* are both very common words. The logarithm is then taken on the ratio to reflect the standard practice of information-theoretic approaches.

Through this definition of mutual information, the relatedness between two words is measured by examining the information content provided by their co-existence -- i.e., by their contribution to the commonality of the documents that contain them. To a certain extent, MI can be considered as a computational means of measuring the commonality of two words (see Section 3.3) in a purely probabilistic sense.

Some desired properties of MI can be observed:

- Association exists between words, => $MI(x, y)>0$
- No interesting relationship, => $MI(x, y)=0$
- Complementary distribution, => $MI(x, y)<0$

Under the corpus-based approach, word relationships are often derived from their co-occurrence distribution in a corpus (Church and Hanks 1989, Hindle 1990, Grefenstette 1992, Myaeng and Li 1992).

Since it is hard to obtain any structural (e.g. hierarchical) relations from corpora, the semantic relations revealed from term associations often represent general, typically associative, relations (Jiang 1996).

4.3.2 Pseudo-Knowledge-based Approach

A natural alternative to strictly distributional techniques for acquisition of semantic association information is the use of manually constructed knowledge bases. With the introduction of machine-readable dictionaries (MRD), lexicons, thesauri, and taxonomies, these manually built pseudo-knowledge bases[6] provide a natural and logical framework for organizing words or concepts into a semantic space. Corpus statistics can then be applied to enhance and fine-tune this qualitative information estimation by providing statistical evidence. Kozima and Furugori (1993) measured word distance by adaptive scaling of a vector space generated from the *Longman Dictionary of Contemporary English* (LDOCE

[6] We call them pseudo-knowledge as they often exhibit the characteristic of a real knowledge base as having a repository containing facts and knowledge, but lack explicit rules about reasoning from the knowledge. Also, many of the constructs were not developed for the purpose of direct machine manipulation.

1987). Morris and Hirst (1991) used Roget's (1977) thesaurus to detect word semantic relationships. With the recently developed lexical taxonomy WordNet (Miller 1990, Miller et al. 1990), many researches have taken advantage of this broad-coverage taxonomy to study word/concept relationships (Resnik 1995, Richardson and Smeaton 1995b).

4.4 Taxonomy and WordNet

WordNet (Miller et al. 1990, Fellbaum 1998) is a large-scale, broad-coverage English lexical taxonomy, manually constructed by George Miller and his colleagues at Princeton University. It consists of several independent semantic networks, each of which corresponds to a part of speech syntax category -- nouns, verbs, adjectives, and adverbs. A word is categorized into the corresponding part of the network based on its syntactic class and semantic sense. The node of the network is called *synset* which represents a set of synonyms. A synset in WordNet represents a lexicalized concept which may contain a phrase or collocation as an element of a synset. For example, one synset/sense for word *tank* is: {*tank, army tank, armored combat vehicle*}. The edge linking two nodes indicates the semantic relationship between the two synsets. There are different relations for each syntactic class network. Synonymy and antonymy relations can be found in all categories. Hyponymy and hypernymy typically exist in noun and verb taxonomies. Also for nouns, there are meronymy and holonymy relations. The most comprehensive and successful hierarchy in WordNet is the Hypernym-Hyponym (IS-A) semantic network for nouns[7].

[7] Strictly speaking, the noun portion of WordNet IS-A taxonomy is not a real tree-like hierarchical network, as occasionally some nodes in the network have multiple inheritances (i.e. more than one parent node). To avoid what might otherwise be inflated calculation of a node's frequency information described in Section 4.5.1, we will consider it as a tree structure in our later algorithms. That is, if one node has more than one super-ordinate node, the first one (the left most one) is always picked as its parent node.

In terms of coverage, WordNet noun taxonomy (version 1.5) contains more than 60,500 synsets, and more than 87,600 entries (index terms). It continues to grow with the latest version, 1.6, released in January 1998.

WordNet is selected as the knowledge source for several reasons. First it is one of the state-of-the-art and most comprehensive lexical ontologies (Noy and Hafner 1997). Second, it is a general taxonomy with no domain dependency. Third, it is built on word semantics and it provides rich semantic association information throughout the network. Last but not least, it is publicly available with continuous upgrades by the developer and support by the research community.

4.5 Semantic Similarity in a Taxonomy

There are certain advantages in the work of semantic association discovery in combining a taxonomy structure with corpus statistics. The incorporation of a manually built pseudo-knowledge base (e.g. thesaurus or taxonomy) may complement the statistical approach where 'true' understanding of the text is unobtainable. By doing this, the statistical model can take advantage of a conceptual space structured by a hand-crafted taxonomy, while providing computational evidence from maneuvering in the conceptual space via distributional analysis of corpora data. In other words, calculating the semantic association can be transformed to the estimation of the conceptual similarity (or distance) between nodes (words or concepts) in the conceptual space generated by the taxonomy. Ideally, this kind of knowledge base should provide reasonably broad-coverage, be well structured, and be easily manipulated in order to derive the desired associative or similarity information.

Since a taxonomy is often represented as a hierarchical structure, (which can be seen as a special case of network structure), evaluating semantic similarity between nodes in the network can make use of the structural information embedded in the network. There are several ways to determine the conceptual similarity of two words in a hierarchical semantic network. Topographically this can be categorized as *node-based* and *edge-based* approaches, which correspond to the information content approach and the graph/conceptual distance approach, respectively.

4.5.1 Node-based (Information Content) Approach

One node-based approach to determining conceptual similarity is called the *information content* approach (Resnik 1992a, 1992b, 1993a, 1995). Given a multidimensional space within which a node represents a unique concept consisting of a certain amount of information, and an edge represents a direct association between two concepts, the similarity between two concepts is the extent to which they share information in common. Considering this in a hierarchical concept/class space, this common information 'carrier' can be identified as a specific concept node that subsumes both of the two in the hierarchy. More precisely, this super-class should be the first class upward in this hierarchy that subsumes both classes. The similarity value is defined as the information content value of this specific super-ordinate class. The value of the information content of a class is then obtained by estimating the probability of occurrence of this class in a large text corpus.

Recall in section 3.3, when following the notation in information theory, the information content (IC) of a concept/class c can be quantified as follows:

$$IC(c) = -\log P(c), \tag{4.1}$$

where $P(c)$ is the probability of encountering an instance of concept c in a corpus. In the case of the hierarchical structure, where a concept in the hierarchy subsumes those lower in the hierarchy, this implies that $P(c)$ is monotonic as one moves up the hierarchy. As the node's probability increases, its information content or its informativeness decreases. If there is a unique top node in the hierarchy, then its probability is 1, hence its information content is 0.

Given the monotonic feature of the information content value, the similarity of two concepts can be formally defined as:

$$sim(c_1,c_2) = \max_{c \in Sup(c_1,c_2)} [IC(c)] = \max_{c \in Sup(c_1,c_2)} [-\log p(c)], \tag{4.2}$$

where $Sup(c_1,c_2)$ is the set of concepts that subsume both c_1 and c_2. To maximize the representativeness, the similarity value is the information content value of the node whose IC value is the largest among those super classes. In another word, this node is the 'lowest upper bound' among those that subsume both c_1 and c_2.

In the case of multiple inheritances, where words can have more than one sense and hence multiple direct super classes, one method to determine word similarity is to select the best similarity value among all the class pairs to which their various senses belong:

$$sim(w_1,w_2) = \max_{c_1 \in sen(w_1)\, c_2 \in sen(w_2)} [sim(c_1,c_2)], \tag{4.3}$$

where $sen(w)$ denotes the set of possible senses for word w.

For the implementation of the information content model, there are some slightly different approaches to calculating the concept/class probabilities in a corpus (Richardson and Smeaton 1995b). Before giving the detailed calculation, we need to define two concept sets: *words*(*c*) and *classes*(*w*). *Words*(*c*) is the set of words subsumed (directly or indirectly) by the class *c*. This can be seen as all words found in each node of a sub-tree in the whole hierarchy, including the sub-tree root *c*. *Classes*(*w*) is defined as the classes in which the word *w* is contained; in other words, it is the set of possible senses that the word *w* has:

$$classes(w) = \{c | w \in words(c)\}. \tag{4.4}$$

Resnik (1995) defined a simple class/concept frequency formula:

$$freq(c) = \sum_{w \in words(c)} freq(w) \,. \tag{4.5}$$

That is, a class *c* frequency is the addition of all the word frequencies in a corpus under the sub-tree where *c* is the root.

Richardson and Smeaton (1995b) proposed a slightly different calculation by considering the word sense factor to reduce the potentially inflated frequency number caused by multiple sense words.:

$$freq(c) = \sum_{w \in words(c)} \frac{freq(w)}{|classes(w)|} \tag{4.6}$$

Finally, the class/concept probability can be computed using maximum likelihood estimation (MLE):

$$P(C) = \frac{freq(c)}{N} \tag{4.7}$$

where N is the frequency of the virtual root class in the WordNet noun hierarchy.

We provide an example to illustrate how node-based approach can be implemented. Assume that we want to determine the similarities between the following classes: (*car, bicycle*) and (*car, fork*). Figure 4.1 depicts the fragment of the WordNet (Version 1.5) noun hierarchy that contains these classes. The number in the bracket of a node indicates the corresponding information content value calculated from a corpus called SemCor which will be used in our later experiment in section 4.8.2[8]. From the figure we find that the similarity between *car* and *bicycle* is the information content value of the class *vehicle*, which has the maximum value among all the classes that subsume both of the two classes, i.e. *sim(car, bicycle)* = 8.30. In contrast, *sim(car, fork)* = 3.53. These results conform to our perception that cars and forks are less similar than cars and bicycles.

[8] As indicated in section 4.8.2, the frequency in SemCor is calculated by class or word sense, not by word. Therefore, there is no need to adjust the class frequency estimation as proposed in formula 4.6.

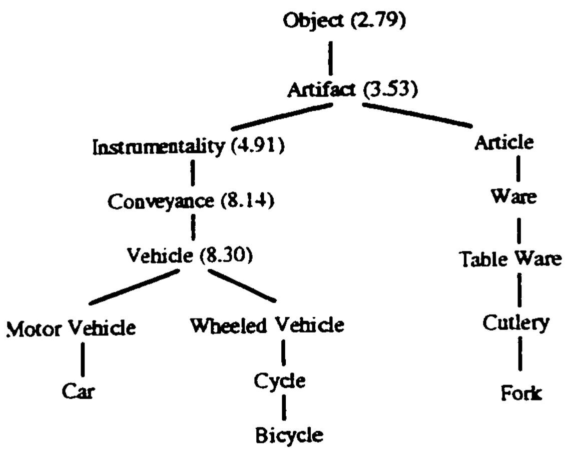


Figure 4.1 Fragment of WordNet noun taxonomy

4.5.2 Edge-based (Distance) Approach

The edge-based approach is a more natural and direct way of evaluating semantic similarity in a taxonomy. It estimates the 'distance' (e.g. edge length) between nodes which correspond to the concepts/classes being compared. Given a network concept space, the conceptual distance can conveniently be measured by the graph distance between the nodes representing the concepts. In doing so for the previous examples using Figure 4.1, the graph distance between *car* and *bicycle* is 5, and between *car* and *fork* is 10. Obviously, the shorter the path from one node to the other, the more similar they are. Similarity is then a decreasing function of the 'distance'.

For a hierarchical taxonomy, Rada et al. (1989) pointed out that the distance should satisfy the properties of a metric, namely: zero property, symmetric property, positive property, and triangular inequality. Furthermore, in an IS-A semantic network, the simplest form of determining the distance between two elemental concept nodes, A and B, is the shortest path that links A and B, *i.e.* the minimum number of edges that separate A and B (Rada et al. 1989).

In a more realistic scenario, the distances between any two adjacent nodes in a path are not necessarily equal. It is therefore necessary to consider that the edge connecting the two nodes should be weighted. To determine the edge weight appropriately, certain aspects should be considered in the implementation. Most of these are typically related to the structural characteristics of a hierarchical network. Some conceivable features are: *local network density* (the number of child links that span out from a parent node), *depth of a node* in the hierarchy, *type of link*, and finally, perhaps the most important of all, the *strength of a single edge link*. We will briefly discuss the concept for each feature:

- *Local Network Density.* With regard to network density, it can be observed that the densities in different parts of the hierarchy are higher than others. For example, in the *plant/flora* section of WordNet the hierarchy is very dense. One parent node can have up to several hundred child nodes. Assuming that the overall semantic mass is of a certain amount for a given node (and its subordinates), the local density effect (Richardson and Smeaton 1995b) would suggest that the greater the density, the closer the distance between the nodes (*i.e.* parent child nodes or sibling nodes).

- *Node Depth.* It can be argued that the distance shrinks as one descends the hierarchy, since differentiation is based on finer and finer details.

- *Link Type.* Type of link can be viewed as the semantic relation type between nodes. In many thesaurus networks the hyponym/hypernym (IS-A) link is the most common concern. Many edge-based models consider only the IS-A link hierarchy (Rada et al. 1989, Lee et al. 1993). In fact, other link types/relations, such as Meronym/Holonym (Part-of, Substance-of), should also be considered as they would have different effects in calculating the edge weight, provided that the data about the type of relation are available.

- *Link Strength.* To differentiate the weights of edges connecting a node and all its child nodes, one needs to consider the link strength of each specific child link. This could be measured by the closeness between a specific child node and its parent node, against those of its siblings. Various methods could be applied here. In particular, this is the place where corpus statistics could contribute. Ideally the method chosen should be both theoretically sound and computationally efficient.

Several studies have been conducted in edge-based similarity/distance determination by responding to the above concerns. Richardson and Smeaton (1995b) considered the first two and the last factors in their edge weight calculation for each link type. Network density is simply counting the number of edges of that type. The link strength is a function of a node's information content value, and those of its siblings and parent nodes. The result of these two operations is then normalized by dividing them by the link depth. Notice that the precise formula of their implementation was not given in the paper.

Sussna (1993) considered the first three factors in his edge weight determination scheme. The weight between two nodes c_1 and c_2 is calculated as follows:

$$wt(c_1,c_2) = \frac{wt(c_1 \rightarrow_r c_2) + wt(c_2 \rightarrow_{r'} c_1)}{2d} \tag{4.8}$$

given

$$wt(x \rightarrow_r y) = \max_r - \frac{\max_r - \min_r}{n_r(x)} \tag{4.9}$$

where $\rightarrow_r$ is a relation of type r, $\rightarrow_{r'}$ is its reverse, *d* is the depth of the deeper one of the two, *max* and *min* are the maximum possible and minimum possible weights for a specific relation type *r*, respectively, and $n_r(x)$ is the number of relations of type *r* leaving node *x*.

Applying this distance formula to a word sense disambiguation task, Sussna (1993) showed an improvement where multiple sense words have been disambiguated by finding the combination of senses from a set of contiguous terms which minimizes total pairwise distance between senses. He found that the performance is robust under a number of perturbations; however, depth factor scaling and restricting the type of link to a strictly hierarchical relation do noticeably impair performance.

Peh and Ng (1997) consider a variant link strength determination called descendant coverage. It calculates the difference in the percentage of descendants by a parent node and that by one of its child nodes. This essentially brings the local density factor into the link strength consideration.

In determining the overall edge-based similarity, most methods just simply sum up all the edge weights along the shortest path. To convert the distance measure to a similarity measure, one may simply subtract the path length from the maximum possible path length (Resnik 1995):

$$sim(w_1, w_2) = 2d_{\max} - [\min_{c_1 \in sen(w_1)\, c_2 \in sen(w_2)} len(c_1, c_2)], \tag{4.10}$$

where $d_{\max}$ is the maximum depth of the taxonomy, and the *len* function is the simple calculation of the shortest path length (*i.e.* weight = 1 for each edge).

4.5.3 Comparison of the Two Approaches

The two approaches target semantic similarity from quite different angles. The edge-based distance method is more intuitive, while the node-based information content approach is more theoretically sound. Both have inherent strengths and weakness.

Rada et al. (1989) applied the distance method to a medical domain, and found that the distance function simulated well human assessments of conceptual distance. However, Richardson and Smeaton (1995b) had concerns that the measure was less accurate than expected when applied to a comparatively broad domain (e.g. WordNet taxonomy). They found that irregular densities of links between concepts result in unexpected conceptual distance outcomes. Also, without causing serious side effects elsewhere, the depth scaling factor does not adjust the overall measure well due to the general structure of the taxonomy (e.g. higher sections tend to be too similar to each other).

In addition, we feel that the distance measure is highly depended upon the subjectively pre-defined network hierarchy. Since the original purpose of the design of the WordNet was not for a direct computation purpose, some local network layer constructions may not be suitable for the direct distance manipulation.

The information content method requires less information on the detailed structure of a taxonomy. It is not sensitive to the problem of varying link types (Resnik 1995). However, it is still dependent on the skeleton structure of the taxonomy. Just because it ignores information on the detailed structure it has its weaknesses. It normally generates a coarse result for the comparison of concepts. In particular, it does not differentiate the similarity values of any pair of concepts in a sub-hierarchy as long as their 'largest common denominator' (i.e. the lowest super-ordinate class) is the same. For example, given the concepts in Figure 4.1, the results of the similarity evaluation between (*bicycle, table ware*) and (*bicycle, fork*) would be the same. Also, other types of link relation information are overlooked here. Additionally, in the calculation of information content, polysemous words will have an exaggerated content value if only word (not its sense) frequency data are used (Richardson and Smeaton 1995b).

4.6 An Integrated Similarity Measure

We propose an integrated approach that is derived from the edge-based principle (hence we keep using the 'distance' notion) by adding the information content as a decision factor. We will consider various concerns of the edge weighting schemes discussed in the previous subsection 4.5.2. In particular, attention is given to the determination of the *link strength* of an edge that links a parent node to a child node.

We first consider the link strength factor. We argue that the strength of a child link is proportional to the conditional probability of encountering an instance of the child concept c_i given an instance of its parent concept p: $P(c_i | p)$.

$$P(c_i | p) = \frac{P(c_i \cap p)}{P(p)} = \frac{P(c_i)}{P(p)} \tag{4.11}$$

Notice that the definition and determination of the information content (see equations 4.1 and 4.5) indicate that c_i is a subset of p when a concept's informativeness is concerned. Following the standard argument of information theory, we define the *link strength* (LS) by taking the negative logarithm of the above probability. We obtain the following formula:

$$LS(c_i, p) = -\log(P(c_i | p)) = IC(c_i) - IC(p)\,. \tag{4.12}$$

This states that the link strength (LS) is simply the difference of the information content values between a child concept and its parent concept.

Considering other factors as we identified in section 4.5.2, such as local network density, node depth, and link type, the overall edge weight (wt) for a child node c and its parent node p can be determined as a multiplication of these factors, together with the just resolved link strength factor, respectively:

$$wt(c,p) = \left(\beta + (1-\beta)\frac{\overline{E}}{E(p)}\right)\left(\frac{d(p)+1}{d(p)}\right)^{\alpha} T(c,p)[IC(c) - IC(p)], \tag{4.13}$$

where $d(p)$ denotes the depth of the node p in the hierarchy, $E(p)$ the number of edges in the child links (i.e. local density), $\overline{E}$ the average density in the whole hierarchy, and $T(c,p)$ the link relation/type factor. The parameters $\alpha\ (\alpha \geq 0)$ and $\beta\ (0 \leq \beta \leq 1)$ control the degree of how much the node depth and density factors contribute to the overall edge weighting computation. For instance, these contributions become less significant when α approaches 0 and β approaches 1.

Although other formulas could be used to control the effects of node depth and edge density, we note that in WordNet (version 1.5) $E(p)$ varies from 0 to 395, $\overline{E}$ is 1.01, and $d(p)$ ranges from 1 to 16, with most values around 7 (see Table 4.2). Thus the chosen combinations of differences, ratios, and exponential introduce a relative measure in a parameterized way which will generate results with relatively small variances. We shall see in Section 4.8 that the formulation we have chosen for measuring edge weight works well in practice.

Total number of nodes (synset)	60,558
Total number of internal nodes	47,110
Total number of nodes that have one child node	4,787
Total number of nodes that have more than ten child nodes	1,043
Average node depth	7.04
Medium node depth	7
Highest node depth	16
Average node density	1.01
Medium node density	0
Highest node density	395

Table 4.2 Some statistics about the WordNet noun taxonomy (version 1.5)

The overall distance between any two nodes in the hierarchy would thus be the summation of edge weights along the shortest path linking two nodes.

$$Dist(w_1, w_2) = \sum_{c \in \{path(c_1, c_2) - LSuper(c_1, c_2)\}} wt(c, parent(c)) \tag{4.14}$$

where $c_1=sen(w_1)$, $c_2=sen(w_2)$, and $path(c_1, c_2)$ is the set that contains all the nodes in the shortest path from c_1 to c_2. One of the elements of the set is $LSuper(c_1,c_2)$, which denotes the lowest super-ordinate of c_1 and c_2. For the shortest path containing n nodes, there are $n-1$ edges. The calculation is done on each edge by walking up from both ends of the path until they meet at a node where their lowest super-ordinate resides. This is why the lowest super-ordinate node is taken away from the path in the calculation. As an example in calculating the weighted distance between *car* and *bicycle* in Figure 4.1, we will sum all the five edge weights:

$$\begin{aligned} Dist(car,bicycle) = {} & wt(car,motor\_vehicle) + wt(motor\_vehicle,vehicle) \\ & + wt(wheeled\_vehicle,vehicle) + wt(cycle,wheeled\_vehicle) \\ & + wt(bicycle,cycle) \end{aligned}$$

In the special case when only link strength is considered in the weighting scheme of equation 4.13, *i.e.* $\alpha = 0$, $\beta = 1$, and $T(c,p) = 1$, the distance function can be simplified as follows:

$$Dist(w_1, w_2) = IC(c_1) + IC(c_2) - 2 \times IC(LSuper(c_1, c_2)) \tag{4.15}$$

An intuitive explanation of this method is as follows. Imagine a special semantic network space where every node (concept) in the space lies on a specific axis and has a mass (based on its informativeness or information content value). The semantic distance between any such two nodes is the difference of their semantic mass if they are on the same axis, or the addition of the two distances calculated from each node to a common node where two axes meet, if the two original nodes are on different axes. It is easy to prove that the proposed distance measure also satisfies the properties of a metric.

For consistency in comparison with other similarity-based measures, we will use measures indicating semantic sameness rather than the difference. Hence our proposed integrated distance measure needs to be converted to a similarity measure. Like the edge counting measure in equation 4.10, the conversion can be made by subtracting the total edge weights from the maximum possible total edge weights. This conversion, as opposed to the conventional one defined in formula 3.19, will ensure linearity between distance and similarity that is often required in empirical evaluations (see section 4.8).

4.7 Semantic Similarity Measures – a Unified View

We now present another perhaps more rigorous and unified view of those semantic distance/similarity measures we have discussed so far in this chapter. We will explain how they can fit into the more general similarity theory we introduced in Chapter 3.

Recall the two basic absolute measures of the relationship between two objects/concepts in section 3.3, when considered from the information-theoretic view of calculating the comparison content. It can easily be observed that the node-based information content measure in formula 4.2 corresponds to the definition of *commonality* of two concepts (formula 3.22), since the lowest superordinate node of any two nodes in the hierarchy inherits the maximal semantic content of the two; while the edge-based formula 4.15, which is derived as a special case of our proposed integrated model, actually corresponds to the definition of *difference* of two concepts (formula 3.23). Since the contents of all the objects/concepts are calculated by the same unit (in information content), it would make sense to have these absolute measures as a type of comparison scheme.

It is then logical to introduce the relative relationship measures here using the similar mapping scheme. For the Dice similarity measure in formula 3.24 we have the corresponding formula 4.16:

$$Dice\_Sim(w_1,w_2) = \frac{2 \times IC(LSuper(c_1,c_2))}{IC(c_1)+IC(c_2)} \qquad (4.16)$$

And for the Jaccard similarity (formula 3.25) we have formula 4.17:

$$Jaccard\_Sim(w_1, w_2) = \frac{IC(LSuper(c_1, c_2))}{IC(c_1) + IC(c_2) - IC(LSuper(c_1, c_2))} \qquad (4.17)$$

For a clearer illustration, Table 4.3 lists the important similarity related concepts, their definitions/derivations, and corresponding semantic content determinations given the knowledge from a taxonomy and the information of concept distributional data from corpora.

Concept	Measure	Semantic Content Determination		
		Name	Calculation	Eq.
Content	$I(A)$	*information content*	$IC(a) = -\log P(a)$	*4.1*
commonality	$comm = I(common(A, B)) = I(A \cap B)$	*node-based*	$IC(LSuper(a, b))$	*4.2*
difference	$diff = I(A) + I(B) - 2 \times comm$	*edge-based*	$IC(a) + IC(b) - 2 \times IC(LSuper(a,b))$ *(special case)*	*4.15*
similarity (Dice)	$Dice\_sim(a,b) = \frac{2 \times comm}{I(A) + I(B)}$	*Dice*	$\frac{2 \times IC(LSuper(a,b))}{IC(a) + IC(b)}$	*4.16*
similarity (Jaccard)	$Jaccard\_sim(a,b) = \frac{comm}{I(A) + I(B) - comm}$	*Jaccard*	$\frac{IC(LSuper(a,b))}{IC(a) + IC(b) - IC(LSuper(a,b))}$	*4.17*

Table 4.3 Similarity related concepts and their semantic content determinations

In this unified view, we have essentially presented a solution to the semantic content determinations for various object comparison schemes that are introduced in our similarity theory described in Chapter 3. We have seen that they all can be fitted appropriately into the general model.

4.8 Evaluation

In this section we proceed to an empirical evaluation of various semantic similarity models that have been discussed. They are Commonality, Similarity (Dice), and Similarity (Jaccard) measures from Table 4.3; Graph Distance (edge counting) measure from formula 4.10; and the proposed integrated Difference approach from formula 4.14. We seek empirical evidence to support our argument that the proposed integrated model can capture well the essence of human similarity judgment. The experiment to be conducted is right on the central problem of lexical semantic similarity: determining the semantic similarity between pairs of lexicalized items such as words or concepts.

4.8.1 Task Description

It would be reasonable to evaluate the performance of various computational measurements of semantic similarity between concepts by comparing them with human ratings on the same setting. The simplest way to implement this is to set up an experiment to rate the similarity of a set of word pairs, and examine the correlation between human judgment and machine calculations. To make our experimental results comparable with other previous experiments, we decided to use the same sample of 30 noun pairs that were selected in an experiment when only human subjects were involved (Miller and Charles 1991), and in another more recent experiment when some computational models were constructed and compared as well (Resnik 1995). In fact, in the Resnik (1995) experiment, he replicated the human judgments on the same set of word pairs that Miller and Charles did. When the correlation between his replication and the one done by Miller and Charles (1991) was calculated, a baseline for human ratings was obtained for evaluation, which represents an upper bound that one could expect from a machine computation on the same task. In our experiment, we compare the proposed integrated model with the node-based information content model developed by Resnik (1995) and the basic edge-based graph distance model, in the context of how well these perform against human ratings (i.e. the upper bound). In addition, we also include two classic similarity measures (Dice and Jaccard) in this experiment.

For the tasks of converting distance-based measures (i.e. the graph distance and 'difference' approaches), we simply use 30 as the value for both maximum graph distance (i.e. 2 times the maximum node depth) and maximum possible total weight. Note this conversion does not affect the result of the evaluation, since a linear transformation of each datum will not change the magnitude of the resulting correlation coefficient, although its sign may change from positive to negative.

4.8.2 Implementation

The noun portion of WordNet (version 1.5) was selected as the taxonomy from which to compute the similarity between concepts. It contains about 60,500 nodes (synsets). The frequencies of concepts were estimated using noun frequencies from a universal semantic concordance SemCor (Miller et al. 1993), a semantically tagged text consisting of 100 passages from the Brown Corpus. Since the tagging scheme was based on the WordNet word sense definition, this enables us to obtain a precise frequency distribution for each node (synset) in the taxonomy. Therefore it avoids potentially spurious results on occasions when only word (not word sense) frequencies are used (Resnik 1995). The downside of using the SemCor data is the relatively small size of the corpus due to the need to manually tag the sense for each word in the corpus. Slightly over 25% of the WordNet noun senses actually appeared in the corpus. Nevertheless, this is the only publicly available WordNet sense tagged corpus that covers complete running texts. The maximum likelihood estimate (MLE) would seem unsuitable for probability estimation from the SemCor corpus. This is because it assigns a zero probability to unseen events in the sample which may not truly reflect the real probability distribution in large data set.

A class of discounting methods has been developed to circumvent the problem of data sparseness by decreasing the probabilities of previously seen events, so that there is a little bit probability mass left over for previously unseen events. Among them, the Good-Turing estimate is a widely used one. Given bin n_r representing the number of items that are observed exactly r times, and k the maximum number of times any item is observed, then

$$\sum_{r=1}^{k} rn_r = N$$

where N is the total size of the sample. To estimate the probability of something that occurs r time, the MLE would simply use the normalized frequency:

$$P_{MLE} = \frac{r}{N}$$

While in Good-Turing method, it first estimates an adjusted frequency r^*:

$$r^* = (r+1)\frac{n_{r+1}}{n_r}$$

It is then normalized to generate the new probability estimate:

$$N^* = \sum_{r=0}^{k} r^* n_r$$

$$P_{GT} = \frac{r^*}{N^*}$$

In our treatment of the SemCor data, we use the Good-Turing estimate with linear interpolation (Resnik 1993a).

4.8.3 Results

Table 4.4 lists the complete results of each similarity rating measure for each word pair. The data on human ratings are from the publication of previous results (Miller and Charles 1991, Resnik 1995). Notice that two values in Resnik's replication are not available, as he dropped two noun pairs in his experiment since the word *woodland* was not yet in the WordNet taxonomy at that time. The correlation values between the similarity ratings and the mean ratings reported by Millers and Charles are listed in Table 4.5. The optimal parameter settings for the proposed similarity approach (i.e. 'difference' measure) are: α=0.5, β=0.3. Table 4.6 lists the results of the correlation values for the proposed approach given a combination of a range of parameter settings. All the correlation coefficients in Tables 4.5 and 4.6 have been tested to be significant beyond the 0.5% level.

Word Pair		M&C means	Replication means	Distance (edge)	Common (node)	Difference (edge)	Sim (Dice)	Sim (Jaccard)
Car	automobile	3.92	3.9	30	10.358	30	1	1
Gem	jewel	3.84	3.5	30	17.034	30	1	1
Journey	voyage	3.84	3.5	29	10.374	27.497	0.836	0.718
Boy	lad	3.76	3.5	29	9.494	25.839	0.766	0.62
coast	shore	3.7	3.5	29	12.223	28.702	0.966	0.933
asylum	madhouse	3.61	3.6	29	15.492	28.138	0.984	0.968
magician	wizard	3.5	3.5	30	14.186	30	1	1
midday	noon	3.42	3.6	30	13.558	30	1	1
furnace	stove	3.11	2.6	23	3.527	17.792	0.214	0.12
food	fruit	3.08	2.1	24	2.795	23.775	0.324	0.193
bird	cock	3.05	2.2	29	9.122	26.303	0.749	0.598
bird	crane	2.97	2.1	27	9.122	24.452	0.652	0.484
tool	implement	2.95	3.4	29	8.84	29.311	0.923	0.857
brother	monk	2.82	2.4	25	2.781	19.969	0.212	0.119
crane	implement	1.68	0.3	26	4.911	19.579	0.368	0.225
lad	brother	1.66	1.2	26	2.781	20.326	0.246	0.14
journey	car	1.16	0.7	0	0	17.649	0	0
monk	oracle	1.1	0.8	23	2.781	18.611	0.187	0.103
cemetery	woodland	0.95	NA	0	0	10.672	0	0
food	rooster	0.89	1.1	18	1.03	17.657	0.092	0.048
coast	hill	0.87	0.7	26	8.917	25.461	0.73	0.575
forest	graveyard	0.84	0.6	0	0	14.52	0	0
shore	woodland	0.63	NA	25	2.795	16.836	0.192	0.106
monk	slave	0.55	0.7	26	2.781	20.887	0.205	0.114
coast	forest	0.42	0.6	24	2.795	15.538	0.187	0.103
lad	wizard	0.42	0.7	26	2.781	20.717	0.23	0.13
chord	smile	0.13	0.1	20	4.452	17.535	0.33	0.198
glass	magician	0.11	0.1	22	1.03	17.098	0.079	0.041
noon	string	0.08	0	0	0	12.987	0	0
rooster	voyage	0.08	0	0	0	12.506	0	0

Table 4.4 Results of Word Pair Semantic Similarity Measurements (30 pairs)

Similarity Method	Correlation (r)
Replication means (human judgment)	0.8848
Distance (edge counting)	0.6004
Commonality (node-based)	0.7941
Difference (edge-based)	0.8282
Similarity (Dice)	0.8177
Similarity (Jaccard)	0.8074

Table 4.5 Correlation between each method and M&C means (30 noun pairs)

Depth Factor (α)	Density Factor (β)			
	β=1.0	β=0.5	β=0.3	β=0.2
α=2	0.79844	0.81104	0.81153	0.80658
α=1	0.80503	0.82255	0.82625	0.82266
α=0.5	0.80874	0.82397	**0.82817**	0.82509
α=0	**0.81127**	0.82284	0.82737	0.82411
α=-1	0.81435	0.81598	0.81818	0.81349
α=-2	0.81315	0.80228	0.80118	0.79492

Table 4.6 Correlation coefficient values of parameter settings for the proposed approach

4.8.4 Discussion

The results of the experiment confirm that the node-based 'commonality' approach proposed by Resnik (1995) provides a significant improvement over the traditional graph distance (edge counting) method. It also shows that our proposed integrated approach (i.e. edge-based 'difference' measure) outperforms the node-based approach. One should recognize that even a small percentage improvement over the existing approaches is of importance since we are nearing the observed upper bound. The results also indicate that two related classical similarity measures (Dice and Jaccard) render a satisfactory performance.

To see if there is really a difference among these methods, we need to test the significance of the differences among their correlation coefficients. Since all these comparison methods are measured on the same 30 pairs of words, the usual Fisher's Z-transformation test would not be valid. Downie and Starry (1977) give a test statistic for comparing two correlation coefficients obtained on the same sample: when we have three variables: 1, 2, and 3, we can compare the correlations of two of the variables with the third. To test whether r_{13} and r_{23} are significantly different they suggest the following statistic:

$$z = \frac{(r_{13} - r_{23})\sqrt{(N-3)(1+r_{12})}}{\sqrt{2(1 - r_{13}^2 - r_{23}^2 - r_{12}^2 + 2r_{13}r_{23}r_{12})}}$$

The above z is interpreted in the usual standard normal distribution when the sample size is large.

With this formula, we are able to do the pair-wise correlation coefficient comparisons for all the computational approaches in Table 4.5, given the human replication account as the third variable. The results confirm that all other methods are significantly better than the graph distance (edge counting) approach at $\alpha = 0.025$ level. However, this test cannot confirm of the significance of differences among these other methods under $\alpha = 0.1$ level. There are two reasons that could possibly explain this, both of which are related to the limitations of this particular test. First, as has been pointed out elsewhere, this test should not be used for small samples (Woods et al 1986:167). In order to compare with others in our word-pair ranking experiment, an important portion of the data that requires human subject input is adopted from previous studies in the literature; we are thus constrained by a size of 30 sample from the original experimental data. Second, there seems to be an implicit assumption for this test statistic that requires exactly three variables. For example, this method tells us to compare two variables by using a third as a reference. If there is more than one variable that could act as this third as a reference, then the result of the comparison could be undetermined. For example, we compare the difference of correlation coefficients between the node-based 'commonality' approach and the edge-based 'difference' approach. If both are referred to the third, the human replication account, there is no significant difference between them. If the third reference variable is selected with the edge-counting method, than the difference would be significant at $\alpha = 0.05$ level. Similar evidence could be found in other pair comparisons.

Since it is difficult to find alternative test statistics for this particular scenario, we acknowledge the results of this test as a principled judgment.

From the unified view we presented in section 4.7, it is noted that the node-based information content measure and the proposed edge-based distance measure correspond to two very different schemes in the theory of object comparison: the 'commonality' and the 'difference' between two objects. The results here seem to indicate, to a certain degree, their expressive power in representing semantic similarity. In particular, it supports the argument that the 'difference' approach may generate a better computational means than that of 'commonality' in determining semantic similarity. To the extent that the 'difference' approach can be used as the basis for a transformation-based model (see Section 3.4), the experiment here will provide empirical evidence supporting the effectiveness of that model.

The results from Table 4.6 conform to our projection that the density factor and the depth factor in the hierarchy do affect (although not significantly) the semantic distance metric. A proper selection of these two factors will enhance the distance estimation. Setting the density factor parameter at β=0.3 seems optimal as most of the resultant values outperform others under a range of depth factor settings. The optimal depth scaling factor α ranges from 0 to 0.5, which indicates it is less influential than the density factor. This would support the Richardson and Smeaton (1995b) argument about the difficulty of the adjustment of the depth scaling factor. Another explanation would be that this factor is already absorbed in the proposed link strength consideration. Overall, there is a small performance improvement (1.69%) over the result when only the link strength factor (α=0, β=1) is considered. Since the results are not very sensitive to the variation in parameter settings, we can conclude that they are not the major determinants of the overall edge weight.

Further examinations of the individual results in Table 4.4 may provide a deeper understanding of the model's performance. The ratings in the table are sorted in descending order based on Miller and Charles (1991) findings. This trend can be observed more or less consistently in six other ratings. However, there are some abnormalities that exist in the results. For example, the pair '*furnace-stove*' was given high similarity values in human ratings, whereas a very low rating (second to the lowest) was found in the proposed distance measure. A further look at their classification in the WordNet hierarchy seems to provide an explanation. Figure 4.2 depicts a portion of WordNet hierarchy that includes all the senses of these two words. We can observe that *furnace* and *stove* are classified under very distinct substructures. Their closest super-ordinate class is *artifact*, which is a very high level abstraction. It would be more reasonable if the substructure containing *furnace* were placed under the class of *device* or *appliance*. If so the distance between *furnace* and *stove* would have been shorter and closer to humans' judgments. This observation re-enforces our earlier thought that the structure of a taxonomy may generate a bias towards a certain distance calculation due to the nature of its classification scheme.

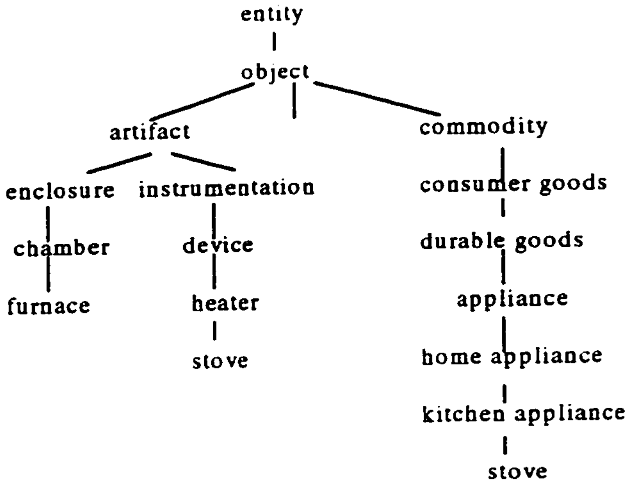


Figure 4.2 Fragment of WordNet taxonomy containing 'furnace' and 'stove'

Table 4.7 shows calculations of the correlation coefficients based on removing the *furnace-stove* pair due to a questionable classification of the concept *furnace* in the taxonomy. The result shows an immediate improvement of all the computational models. In particular, our proposed integrated model ('difference' measure) indicates a large marginal lead to the node-based 'commonality' measure.

Similarity Method	**Correlation (r)**
Distance (edge counting)	0.6042
Commonality (node-based)	0.8191
Difference (edge-based)	0.8654
Similarity (Dice)	0.8505
Similarity (Jaccard)	0.8430

Table 4.7 Correlation between each method and M&C means (29 noun pairs, removing the '*furnace - stove*' pair)

Whether the apparent extreme value for the 'furnace-stove' pair is an outlier or just happens by chance cannot be determined at this moment. This idiosyncrasy would perhaps best be left to the later large scale test (in chapter 5) when both the size of samples is large and adequate statistical tests are available.

4.9 Related Work

Closely related works to this study are those that were aligned with the thread of our discussion. In the line of the edge-based approach, Rada et al. (1989) and Lee et al. (1993) derived semantic distance formulae using the edge counting principle, which were then used to support higher-level result ranking in document retrieval. Sussna (1993) defined a similarity measure that takes into account taxonomy structure information. Resnik's (1995) information content measure is a typical representative of the node-based approach. More recently, Richardson and Smeaton (1995b) and Smeaton and Quigley (1996) also worked on an integrated approach of combining both node-based and edge-based approaches. The differences between theirs and ours are determinations of elements and methods in the integration process. In particular, we see our link strength determination as a more theoretically sound method.

As will be explored in the next chapter, one of the many applications of semantic similarity models is for word sense disambiguation (WSD). Agirre and Rigau (1995) proposed an interesting conceptual density concept for WSD. Given WordNet as the structured hierarchical network, the conceptual density for a sense of a word is proportional to the number of contextual words that appear in a sub-hierarchy of WordNet where that particular sense exists. The correct sense can be identified as the one that has the highest density value.

Using an online dictionary, Niwa and Nitta (1994) built a reference network of words where a word as a node in the network is connected to other words that are its definitional words. The network is used to measure the conceptual distance between words. A word vector is defined as the list of distances from a word to a certain set of selected words. These selected words are not necessarily its definitional words, but rather certain types of representational words called *origins*. Word similarity can then be computed by means of their distance vectors. They compared this proposed dictionary-based distance vector method with a corpus-based co-occurrence vector method for WSD and found the latter has a higher-precision performance. However, in a test of leaning positive or negative meanings from example words, the former gave remarkably higher precision than the latter. Kozima and Furugori (1993) also proposed a word similarity measure by spreading activation on a semantic net composed by the online dictionary LDOCE.

In the area of IR using NLP, approaches have been pursued to take advantage of statistical term association results (Strzalkowski and Vauthey 1992, Grefenstette 1992). Typically, the text is first parsed to generate syntactic constructs. Then the *head-modifier* pairs are identified for various syntactical structure. Finally, a specific term association algorithm (similar to the mutual information principle) is applied to the comparison process on a single term/concept basis. Although only modest improvement has been shown, the significance of this approach is that it does not require any domain-specific knowledge or the sophisticated NLP techniques. In essence, our proposed combination model is similar to this approach, except that we also resort to extra knowledge sources—machine readable lexical taxonomies.

4.10 Summary

In this chapter, we have presented a new approach for measuring semantic similarity between lexicalized words and concepts. It combines the lexical taxonomy structure with corpus statistical information so that the semantic distance between nodes in the semantic space constructed by the taxonomy can be better quantified with the computational evidence derived from distributional analysis of corpus data. Specifically, the proposed measure is an integrated approach that inherits the edge-based approach of the edge counting scheme, enhanced by the node-based approach of semantic content calculation.

When tested on a common data set of word pair similarity ratings, the proposed approach outperforms other computational models. It gives the highest correlation value (r=0.828), with a benchmark resulting from human similarity judgments, whereas an upper bound (r=0.885) is observed when human subjects are replicating the same task.

In our approach to resolving lexical semantic similarity using taxonomy knowledge, we essentially secured a solution to determining the 'commonality' of two objects in information-theoretic terms. This is crucial to a realization of computational means of resolving universal lexical semantic similarity.

Compared with other relevant models in the literature, the proposed model has certain advantages. It is theoretically sound as it can be translated to a comparison scheme in object comparison theory. It has a practical appeal as it is easy to implement without requiring a complex and sophisticated knowledge base. Finally, it is domain independent.

In testing our developed model, we replicated the experimental conditions and data conducted from the literature. However, this may still not be sufficient to support our arguments as the sample base of the original 30 noun pairs appears to be too small. In the next chapter, we will apply all the models tested in this chapter to a much larger domain of an NLP task, the lexical semantic disambiguation. We will then see how these models perform in that test.

Chapter 5

Semantic Similarity and Word Sense Disambiguation

5.1 Introduction

A further application of the proposed semantic similarity model is on the lexical level of an NLP task: Word Sense Disambiguation (WSD). In this chapter, we propose a simple approach to the WSD task. This method belongs to the unsupervised model which does not need any previously trained (sense tagged) data. As in many WSD methods, we intend to use maximally the cues from the local contextual information of the target word. The main purpose for performing this WSD task is to demonstrate the effectiveness of such a simple method, and particularly to seek empirical evidence which would further support the claim that the proposed similarity model can generate better performance than other general similarity models in NLP applications. To accomplish this, a WSD experiment is conducted to compare the proposed similarity model with other computational accounts that are used in the experiment of the last chapter. The results of the experiment further verify the improvement of the proposed model over other models given a task that covers a large sample data set to generate sufficient statistical significance.

5.2 Background about WSD

5.2.1 Defining the WSD Problem

Word Sense Disambiguation (WSD) is a task of determining which of the senses of an ambiguous word is involved in a particular context that belongs to the definition of a dictionary or lexicon. Like parsing or part-of-speech tagging, it is a typical 'intermediate' NLP task which serves the purpose of providing intermediate results for those higher level 'final' tasks such as Information Retrieval (IR) or Machine Translation (MT). Also by 'intermediate' we mean that the evaluation of the task is determined by some linguistic or theoretical criterion, as opposed to those 'final' tasks where the criteria can be judged by end users (Wilks and Stevenson 1996). The nature of such intermediate tasks makes them more versatile and less independent so that a standardized performance evaluation scheme is difficult to derive. Nevertheless, this kind of 'relativity' has certainly led to a rich and prominent research field in NLP.

5.2.2 The Challenge of the WSD Task

Compared with other intermediate tasks, the degree of difficulty for WSD is much higher, as it involves not only the syntactic information of the language but a knowledge of contextual semantics as well. In addition, as pointed out in Ng (1997), Resnik and Yarowsky (1997), and Wilks and Stevenson (1996), there are other factors that add to the difficulty of the WSD problem:

- No consensus of the definition of word sense. The division of word meanings into distinct dictionary or lexicon senses is frequently arbitrary. There is rarely a mapping in terms of set inclusion between the differing sense sets for a word in different dictionaries.

- Granularity of classified senses. Depending on different lexicographers, the resultant sense definitions range from a very coarse 'cluster' like classification, to a very fine-grained sense classification. Dealing with the distinction of homographs and sense also aggregates the difficulty.

- Knowledge acquisition bottleneck. There is a lack of adequate MRDs, lexicons, thesauri that record relevant sense information. Even if the information is there, it is frequently hard to digest. Also, adequately large sense tagged data sets are difficult to obtain due to the high cost of effort required by human annotations to corpora.

- Evaluation systems not yet standardized. This is essentially the result of the above mentioned problems, as researchers have tended to use different sets of testing words, different corpora, and different evaluation metrics.

5.3 Approaches to WSD

There are basically two approaches to the WSD problem based primarily on the source of information used: corpus-based statistical methods and knowledge-based methods. We briefly discuss the basic ideas of each.

The statistical approach uses a training corpus to construct, for each sense of a target word, the distribution of closed-class words, parts of speech, and open-class words found within its immediate context. A classifier program then uses statistical learning models (e.g. the Bayesian algorithm) to select the most likely sense of each new occurrence based on these training distributions.

Typical knowledge-based approaches use MRDs, lexicons, thesauri and other knowledge sources to generalize from training examples to new instances. The semantic context of each sense of a target word is represented by its syntactical contexts from the training corpus. For each new occurrence, this contextual information is compared with the information in the sense definitions in the knowledge base, and the sense that is the most similar is selected.

There are also many models that combine these two approaches to take advantage of information from both corpus statistics and knowledge bases.

5.3.1 Sources of Information for WSD

There are several sources of information a WSD method usually employs. They mainly consist of internally existing contextual information from the text itself, collocation information from corpora, and high-level semantics from external knowledge sources.

- Syntactic context. This is the basic and commonly used source for the WSD problem as the sense of a word is essentially defined by its context. Typical methods use a window of words surrounding the target word (Miller et al. 1994, Leacock et al. 1993, and Yarowsky 1992). The length of the window could be fixed by a certain number of words or could be varying depending on some other syntactic structure (e.g. sentence boundary).

- Collocations. This method assumes that a word occurring in various similar (syntactic) contexts would tend to preserve a similar sense. Most statistical approaches utilize this important information. Dorr (1996) uses it for verb WSD problem. Yarowsky (1993) developed a model based on the notion of "one sense per collocation." In contrast to the major views of local collocations, Lin (1997) stated that different words would likely have similar meanings if they occur in identical local contexts. He employs this intuition for his WSD task.

- Part-of-speech tags. Grammatical annotated information can often provide value towards a solution of a proper sense determination (Bruce and Wiebe 1994, Wilks and Stevenson 1996). In fact, a generated text tagged with parts-of-speech is often a pre-requisite for many WSD models (see Richardson and Smeaton 1995b, Ng and Lee 1996, Dagan et al. 1997, and Wilks and Stevenson 1997).

- Selectional preference and word-class semantics. Resnik (1997) explores a statistical model for capturing the co-occurrence behavior of predicates and conceptual classes in a taxonomy. The relationship between selectional preference acquisition and WSD seems a circular one. McCarthy (1997) describes a model of applying WSD for acquisition of selectional preference.

- Dictionary/lexicon definition. The simplest use of dictionary data is computing the overlap of words between the dictionary definition and the sentence containing the target word (Lesk 1986). Furthermore, Luk (1995) uses the Longman dictionary (LDOCE 1987) to manually build a list of 1792 defining concepts which act as a set of controlled vocabulary to define each entry in the dictionary. The co-occurrence information of each pair of the defining concepts is then used to help determine the sense of a target word. Translations of senses in a bilingual dictionary is another source of information for the WSD task (Dagan and Itai 1994).

- Thesaural categories. In Yarowsky's experiment (Yarowsky 1992), Roget's Thesaurus (1977) categories serve as approximations of conceptual classes. The strategy is based on the observation that different conceptual classes of words tend to appear in recognizably different contexts, and different word senses tend to belong to different conceptual classes. Therefore if one can differentiate the conceptual classes, one would then effectively differentiate the word senses that are members of those classes.

Many recent WSD models employ several of the above-mentioned resources when a combined corpus and knowledge base approach is adopted (Ng and Lee 1996, Rigau et al 1997, Wilks and Stevenson 1997).

5.3.2 Performance Evaluation

The simple and almost exclusively used evaluation criterion for the WSD problem is the exact match or simple accuracy criterion. It is the percentage of the number of correctly tagged sense tags over all assigned sense tags.

The range of such a performance value for a typical WSD task can often be determined, although it is not a common practice in many WSD experiments. The upper bound for WSD is usually human performance as we should not expect an automatic procedure to do any better if human judges disagree on the correct assignment for a particular context. For the lower bound, there are typically two considerations. The first uses random choice as a baseline, where a sense for a particular word is selected randomly among all the possible alternatives (Miller et al 1994, Resnik 1997). The second method always picks the most frequent sense for a word for all the contexts. This is a simple yet very effective heuristic as it often surpasses many elaborate WSD algorithms for certain tasks. A simple look-up in a dictionary may obtain this information since most desktop dictionaries list senses in decreasing order of frequency.

The actual performance data for specific WSD algorithms vary depending on the test set, the corpora, and particularly the chosen granularity of senses. For a small test set and very coarse sense (e.g. typical homograph) distinctions, the achievements have been very high. Yarowsky (1992) achieved a 92% accuracy for a dozen words when the system was trained on a 10 million word Grolier's Encyclopedia. In Luk's (1995) algorithm, he tested the same 12 words, but used the much smaller Brown corpus of a very different genre. The result was an average of 77% accuracy. Interestingly, this result is 6% higher than human judgment when the same condition was applied. For a test that targets six binary sense words, Gale et al. (1993) achieved an average accuracy of 92%.

For the more refined-sense test data, it is not surprising that the performance of WSD algorithms would be much lower. This is more evident when the test set is large or uses the whole running text instead of a small, carefully selected set of words. With the very fine-grained WordNet sense definitions and 191 words to test, Ng and Lee (1996) reported a 54% accuracy for 50 files of the Brown corpus, and 68.6% accuracy for the Wall Street Journal test files. With the same WordNet sense definition, Lin (1997) used 7 complete running files from SemCor as a test set, and achieved a 56.1% accuracy.

5.4 Applying Semantic Similarity to WSD

To reiterate, our main purpose for conducting this WSD experiment is to make a comparison between the proposed similarity model and other related similarity models when all are applied to the same NLP application. Therefore, rather than constructing a sophisticated WSD algorithm, we intend to provide an environment to minimize the interference from other contributing factors for such a comparison. Hence the WSD algorithm we will employ is a rather simple one that focuses on utilizing one main information source in the disambiguation process.

Our approach is essentially a knowledge-based approach with the aid of conceptual knowledge from the WordNet taxonomy to help determine the sense for an ambiguous word in a context. It is a simplified knowledge-based approach as it requires no sense-tagged training data nor any other corpora for training purposes.

The source of information mainly used in this algorithm is one of the features called *locality* (Richardson and Smeaton 1995a) or *surrounding words* (Ng and Zelle 1997). Specifically, it contains a window of unordered content-bearing words (i.e. nouns) surrounding the target word in the text. The other implicit knowledge source is a lexicon (WordNet) through which the semantic relations between contextual words and the target word can be explored. We do not employ any training data nor any pre-tagged sense data.

The essential algorithm for this method is to calculate the semantic similarity between each sense of the target word and its contextual words, and pick the sense that has the overall maximal similarity value. The method is similar to the mutual constraints from Sussna

(1993) and locality disambiguator from Richardson and Smeaton (1995a). One justification for this is that senses of all the contextual words tend to form a coherent sub-space in the overall semantic space. This bears a resemblance to the operation of seeking a maximal *conceptual density* in the work by Agirre and Rigau (1995). Another argument is that a highly semantically similar word in the context will trigger the confirmation of the proper sense of the target word by identifying the shortest path that links this contextual word and the target word in the semantic space. This is supported by the process of creating a lexical chain from contextual words which acts as a semantic representation of the context that can lead to a clarification of the ambiguity of the target word (Hirst and St-Onge 1998).

5.5 Evaluation

In this section we will conduct experiments to test the proposed similarity-based WSD method. In particular, we will compare the performance of various similarity models under this same 'intermediate' NLP task. All the lexical similarity methods discussed and tested in Chapter 4 are repeated here for comparison. For simplicity, the simplified format (formula 4.15) of 'difference' approach is employed here.

5.5.1 Task Description

To make it more practical our WSD task will tag every content word (i.e. nouns) in an open running text. The input consists of the original stream of sentences of the text. In the output, each content word is given a tagged sense number based on the WordNet sense definition.

As the only publicly available sense tagged corpus with WordNet sense definitions, SemCor corpus is selected as the test set. In fact, we use the "press report" part of the SemCor corpus, which contain 7 files with about 2000 words each.

Since all the similarity models use only nouns in the WordNet as a source of information, we first need to filter the corpus by picking out all the nouns from the test set. Proper nouns are neglected since they usually are not classified in WordNet. The result is a total occurrence of 2,907 nouns, of which 2,089 are polysemous.

5.5 2 Results and Discussion

We have tested various lengths of the moving window as the contextual information and found that the results are not very sensitive to the length of the window. Table 5.1 reports the results for the length of window as 4, i.e. there are 2 words (nouns) on each side of a target word to be used in this sense disambiguation process. The second column lists the count of correctly tagged polysemous words for each method. The corresponding accuracy is calculated in the third column (i.e. the number in the second column divided by 2,089). The overall accuracy including both monosemous and polysemous words (a total of 2,907) is provided in the last column.

Method	Correct Answers for Polysemous Words	Precision (Polysemous Words)	Precision (Overall)
Random Sense Selection	572	27.4%	47.7%
Most Frequent Sense	1438	68.8%	77.6%
Node Based ('commonality')	898	43.0%	59.0%
Edge Based (edge counting)	845	40.4%	57.2%
Edge Based ('difference')	1180	56.5%	68.7%
Similarity (Dice)	952	45.6%	60.9%
Similarity (Jaccard)	935	44.8%	60.3%

Table 5.1 Summary of Results for WSD

With the Z-test for the equality between two proportions, we are able to confirm that all the computational models are significantly better than the random sense selection baseline. However, they are all significantly weaker than the most frequent sense heuristic. This is hardly surprising, since the frequency statistics are obtained from the whole SemCor corpus of which the test data is a subset. The fact that the most frequent sense is used in this subset 68.8% of the time merely reflects the skewness of the frequency data. Nevertheless, it is interesting that trying to improve on this by using local contexts leads to a deterioration instead.

The best computational model (the 'difference' approach) is still 8.9%–12.3% behind this simple heuristic. There are several reasons that may explain why: First, given the main purpose of this WSD application is to compare the performance of various similarity methods, we try to eliminate effects that may otherwise come from other contributing factors by limiting the sources of information to be employed. We only use surrounding words as the major source of information. Second, the information is not fully utilized even with this local contextual source, as the words are unordered and only nouns are selected (hence a loss of both syntactic and collocational information). Third, WordNet has a very fine-grained sense definition, whereas in many cases more than one sense can apply to a certain scenario (Richardson and Smeaton 1995a). Fourth, human errors are also high in manual sense assignments as is evidenced in the original manual sense tagging of the SemCor corpus.

In fact, the task we have chosen is indeed a very challenging one due to the fine-grained WordNet senses and unrestricted targets in a running text. This can be better demonstrated when we compare it with other research results when similar tasks are performed. We list other results that are based on the SemCor corpus that our task used. In an unsupervised WSD algorithm using mainly the selectional constraints information, Resnik (1997) achieved accuracy in the range of 35.3% and 44.3% for ambiguous nouns. Agirre and Rigau (1995) reported an accuracy of 47.3% with their conceptual density formula. The closest comparable task is Lin's (1997), when the same 7 files in the SemCor were chosen for the test. He used syntactic information from a large corpus to aid sense determination and achieved an accuracy of 56.1% for polysemous nouns, and the corresponding score for the most frequent heuristic is 58.9%. Notice that there is a difference in the target words selection as Lin's targets also included proper nouns (as 3-way ambiguous words: person, organization, or location). Hence his baseline figure is different from ours. Based on the above references, we can observe that our simple algorithm is capable of generating comparable or superior results.

When the results among the five various similarity based methods are examined, we can draw several conclusions. First, the relative performance for each method is in line with that found in previous tests when all are applied to a word-pair similarity ranking experiment conducted in Chapter 4. Second, results from this relatively large sample space test further confirm the conclusion drawn there that our proposed integrated model outperforms other related computational methods. Third, to see if the performance is domain or size dependent, we did a cross-data examination. In testing using a different corpus (MUC-4 terrorism domain), for the polysemous words Peh and Ng (1997) obtained an accuracy of 39.8% with the information content (i.e. the node-based 'commonality') method, and the corresponding most frequent heuristic is 63.2%. This range of results is very close to what we have obtained here. Therefore, we are confident about our selection of the test sample size and domain. Interestingly, the integrated ('difference') approach yields significantly better results than the other similarity methods. The differences are all significant at 99.9% confidence level in a *Z*-test for two proportions. As we have argued earlier, this method resembles the essence of the transformation-based models. To the extent that the 'difference' approach can be used as the basis for a transformation-based model, the experiment here will provide empirical evidence in supporting the effectiveness of the transformation-based model.

Chapter 6

Semantic Similarity and Business Catalog Retrieval

6.1 Introduction

To raise the complexity of the issues surrounding semantic similarity we move from conducting a single, elementary concept similarity comparison (in Chapter 4) to a multi-layered compound concept (phrase-like) similarity comparison. A final and more practical application of the proposed similarity model is in the areas of text retrieval and document classification. This actually was the motivation for the research work that was targeting building an Electronic Industrial Directory (EID) — a Yellow Pages-like online business directory for storing and retrieving trade-related business information (see Appendix B). The main function of the EID is to match the product/service descriptions provided by a purchaser and a supplier so that a potential business exchange could be further pursued.

One of the decomposed tasks of the EID functions is to organize the business directory in a fashion that each company's product/service description is properly classified under a particular business category. The Standard Industrial Classification (SIC), which covers a wide range of business category information, is chosen to represent the framework of the EID database. Therefore a function is needed, for both the business classification and subsequent query retrieval, to locate the most relevant SIC heading(s) under which a new product/service description would belong. For the SIC headings match, we develop algorithms to parse both SIC headings and the incoming query as short product/service descriptions, to compare their corresponding decomposed parts, and to arrive at an overall semantic similarity score for each candidate heading. The semantic similarity scheme developed in Chapters 3 and 4 is applied at various stages of the matching process.

This chapter describes the implementation details and performance evaluation of the SIC headings search and retrieval process. First, since the pattern of an SIC heading can be generalized as a compound concept, the principles of comparing complex concepts are discussed in terms of a general similarity model. Second, complex concepts (given SIC headings as specific instantiations) analysis and retrieval are described. For the analysis, a parsing algorithm is developed to tackle the syntactic ambiguity problems that are typically encountered within a sub-sentence language structure, namely, noun compounding, prepositional phrase attachment, and co-ordination and conjunction. Third, a sample data set from a real world business application is used to test the implemented SIC headings search prototype. Various similarity model schemes are evaluated against the benchmark of the classical vector space model.

6.2 Principles of Complex Concept Comparisons

For complex concepts (represented in compound phrases) similarity comparisons, we take advantage of the classic feature-based models, with consideration of possible feature alignment. We treat the whole complex concept as a single, though complicated, concept/object with many features derived from various parts of the compound phrases. A decomposition step is required in order to understand and match the corresponding subcomponents/features. This essentially corresponds to the feature alignment process. Once features are discerned and appropriately aligned, similarity between lower level lexicalized items can be calculated using previously defined approaches. Finally, the component similarities are aggregated into an overall similarity measurement, with certain weighting schemes. This methodology reflects the principle of the general similarity measure defined in formula 2.17. In particular, feature similarity is not a simple features overlapping determination, but rather a recursive application of lower level similarity calculations.

According to the EID design principles outlined in Appendix B, in both SIC headings analysis (parsing) and the retrieval prototype development, we intend to stay as much domain independent as possible, so that the developed models can have a wider applicability to tasks in other domains. In the parsing process, rather than conducting a complete syntactic analysis with some polished techniques, we employ some simple yet effective partial parsing techniques which are sufficient to generate structural regularities whereby semantic similarity models can be further applied. These so called *weak techniques* (Grefenstette 1994) require no pre-existing domain-specific knowledge structures, yet are able to recognize enough structural regularities in text that we need.

6.3 Implementation of the SIC Headings Retrieval

As discussed in the EID design in Appendix B, the two procedures for new business category classification and query retrieval both require the step to fetch some SIC heading(s) that are relevant to the user's entry. As for the implementation, this step can be decomposed into two major procedures: a) heading/entry parsing, and b) headings search and retrieval.

In this section, we will present a prototype implementation of the SIC headings retrieval subsystem. We discuss in detail the two main steps for this function. First, for the headings parsing we will focus on discussing the SIC headings analysis. As will be evidenced in the discussion in section 6.4.5, users' entries are typically short and less complicated in structural representation, which can often be treated by the same SIC headings analysis scheme. Secondly, we will present a retrieval algorithm that is to fetch and rank the relevant headings.

6.3.1 The SIC Headings Analysis

Unlike complete sentence structure in a typical natural language text, the SIC headings are typically short in a phrase-like structure. Also, they do not exhibit rich linguistic features, rather a simple statement about objects and facts. However, as typical of any natural language text, the analysis of SIC headings still faces resolving several potential syntactic ambiguities. The common ones are: noun compounding, prepositional phrase attachment, co-ordination and conjunction.

6.3.1.1 The grammar of the SIC headings

A study of the headings used in the SIC system and the patterns submitted by users to address a business category indicates that this field is usually expressed as a single noun, a compound noun, or several compound noun phrases separated by some punctuation marks (e.g. ',', ':', or '-'). This syntactical character often reveals its semantic indications. The noun or compound noun before the punctuation mark (if any) often indicates the core concept in the query or heading, whereas any further noun phrase following the punctuation mark expresses a further description/constraint to the core concept, as well as indicates the industry type if necessary. Thus we can roughly express a full length business description (typically in the pattern of a SIC subheading) in the following sequence:

<core part> [<punctuation><descriptors>] [<punctuation><constraints>] [<punctuation><industry type>]

where the <core part> has to be present and the rest is optional. The example (1) shows a heading with all four parts present.

```
(1)   3949 Ammunition belts, sporting type: of all materials—mfg
       ↑        ↑               ↑             ↑           ↑
      code     core         descriptor    constraint  industry type
```

A further observation and analysis of SIC headings shows that they are better structured than a typical user's description. This seems obvious and logical. But still much effort has been expended in order to generalize a rule set to parse the headings. Lower-level headings (mostly subdivisions under level four), which is our main target of interest, can usually be represented in the syntax outlined in Figure 6.1 (given in BNF).

```
          <SIC heading>  ::= <core part> [<descriptor part>] [<constraint part>]
                             [<industry type part>].
             <core part>  ::= <terms> [ “(“ <core annotation> “)”].
       <core annotation>  ::= <terms>.
       <descriptor part>  ::= “,” <descriptor> [ “(“ <descriptor annotation> “)” ].
            <descriptor>  ::= <terms>.
 <descriptor annotation>  ::= <terms>.
            <constraint>  ::= “:” <terms>.
    <industry type part>  ::= “—“ <industry type>
         <industry type>  ::= “mfg” | “contractors” | “government” |... | “wholesale”.
                 <terms>  ::= <terms> <term>
```

Figure 6.1 A general syntax for SIC headings

This set of syntax rules mainly describes the parsing of a heading when a punctuation mark is encountered. More-specific rules under a parsed subcomponent are omitted here as the situation becomes complicated and varies depending on what corresponding part resides in a heading. This is why the last entry <terms> is defined recursively as a concatenation of any bunch of <term>, where <term> is a terminal item that can be any word. Further treatment of <terms> will be detailed later.

As the original design of SIC headings was intended to assist experts/agents to classify businesses, the grammatical rules we derived from the observation of the heading patterns are, hence, not that rigorous. There are many variances and exceptions to our abstraction. Nevertheless, in general the SIC headings are much more structured than their counterparts—user defined business categories.

As we have just stated, each subcomponent part can be decomposed into a set of words or terms, each representing a single unique concept. Typically, a word-level representation can be categorized into one of the five major types in terms of conceptual level functional decomposition:

- *Core term*, identifying the single main subject of an industry/business class. It can refer to an object, e.g. *rims*, *telephone_booth*[9], *syrup*; or names of professions, e.g. *surgeon*, *investor*;

- *Descriptor/modifier/constraint*, describing the characteristics of the core subject in the query or heading. This is usually one (or more) noun or adjective that is prefixed to the 'core term' in the specification where a constraint is specified. For example, the terms *car wheel* in the compound *car wheel rims*, the adjective *underwater* in the phrase *underwater telephone*;

- *Predicate*, expressing the 'action' or activity of an establishment, e.g. *manufacturing*, *repair*, etc. Mostly they are in the form of a gerund, as the word *cleaning* in the heading *Air cleaning systems*;

- *Connector*, it can be a conjunction that indicates a conjoined part, or a preposition that lead to a preposition phrase attached to the main part; for example, the conjunction word *and* in the phrase *dog and cat food*, and the preposition *for* in the heading *Greenhouses for food crops*;

- *'Empty' words*, those closed-class morphemes that contribute less to the significance of the concepts represented in the description, such as *more*, *as*, *some*, etc. This corresponds to the 'stop-word' concept in an IR expression.

As we shall see shortly, from the above categories, those closed-class morphemes (e.g. a *connector*) or those with special morphological characteristic (e.g. a *predicate*), together

[9] A collocation is often treated as one single term.

with the punctuation mark, can often provide significant clues to the determination of the boundary of a specific subcomponent in the parsing process.

6.3.1.2 User's Query Analysis

As a comparison, we present a brief analysis of the counterpart of the SIC headings—the representation of a user's query. An observation of the patterns of user defined industry categories[10] indicates that they are usually very specific in content expression, which tends to correspond to a lower-level classification in the SIC hierarchy. This would help identify headings mostly from the fourth level (i.e. *industry classes*) of the hierarchy and their subheadings to obtain a better precision value for retrieval because of the narrowed search scope. Another characteristic is that they are often very concise and mostly contain only the 'core part' format in the content representation. Table 6.1 lists a sample of user's inputs and corresponding SIC headings. In general, we can have similar treatment for a user's query as we do for SIC headings. From now on, we will focus on how to further analyze the 'core part' as a compound noun or phrase and decompose it into a set of atomized concepts so the similarity comparison can be conducted.

User's Query Input	Corresponding SIC headings
Canned Chickens	2015 Chickens, processed: fresh, frozen, canned, or cooked--mfg 2015 Poultry, processed: fresh, frozen, canned, or cooked--mfg
Sports Apparel	2329 Sports clothing, non tailored: men's and boys'--mfpm—mfg 2329 Athletic clothing: men's and boys,--mfpm—mfg 5136 Sportswear, men's and boys'--wholesale 5137 Sportswear: women's and children's--wholesale
Diluents/Thinners	2851 Thinner, lacquer—mfg 2851 Thinners, paint: prepared—mfg 2869 Solvents, organic—mfg

Table 6.1 A sample of user's inputs and corresponding SIC headings

[10] A complete list of queries chosen for the prototype evaluation can be found in Appendix A.

6.3.1.3 Subcomponent Analysis

We have noted that 'core part' is a crucial subcomponent for a successful retrieval as it contains the main subject information about a heading. In addition, the structure of the 'core' part is often more complex than the other three parts we identified above. Several examples of SIC headings whose 'core part' contains the word *steel* are presented here to illustrate the possible structures of a 'core part'. These are listed with the increasing degree of complexity for a 'core part'.

- Single Noun: *Steel*
- Noun Compound: *Razor blade strip steel*
- PP Attachment: *Shot peening-treating steel to reduce fatigue*
- Compound Phrases with Conjunction: *Steel tire cord and tire cord fabrics*

From the above examples we can see that the 'core part' alone in the headings is already much more complicated than a single, atomized concept normally expressed in one word that we have encountered in the early chapters. This is still not a full sentence though. It just represents a complex concept with other concepts modifying or restricting the central concept/subject which we call the *core term*. For each category listed above, we will discuss some of the strategies that tackle the complexity of parsing the 'core part', in particular, the ways to correctly identify the *core term* in that part. To keep our solution as simple and domain independent as possible, we intend to keep using syntactic rules as much as possible, and to resort to semantic analysis only when the need arises. Even with grammatical rules, instead of full text parsing, we employ some 'shallow' parsing schemes that are simple yet effective to recognize the structural regularities that we need. We rely heavily on discerning those constituents (e.g. closed-class morphemes) that are evident in indicating the structure of a 'core part'.

Single Noun Analysis. The process of this first scenario is trivial, as the 'core term' is equal to the whole 'core part'. This situation does not happen very frequently; it only represents about 5% of the whole SIC headings list (under the 4 digit level).

Noun Compounds Analysis. The second scenario falls in the domain of lexical semantics in classical linguistic study and NLP research. This particular linguistic construct, we refer to it as noun compounds, has received regular attention from linguists. There are even many names for it: ***noun compounds, compound nouns, nominal compounds, complex nominals, noun-noun compounds, noun sequences***, etc. While the definitions vary, all these terms describe very similar classes of linguistic constructions. All include at least a noun, and all involve multiple open-class morphemes; whereas members of closed-class morphemes are typically not involved. The difference among them is typically the degree of allowing for certain types of open-class morphemes to be included. The most restrictive definitions use the term 'noun-noun compounds' (Downing 1977) or the term 'noun sequences' (Leonard 1984), which only allow for a sequence of nouns to be considered. Lauer's (1995b) 'noun compounds' definition relaxes things a bit to allow for gerunds to be included. A further relaxation by Levi (1978) named it as 'complex nominals,' which include certain non-predicating adjectives before the noun (e.g. *electrical engineer*). The most open definition is called 'noun premodifiers' (Quirk et al 1985) which permit virtually any constituent to appear before a noun to form a pre-modified noun.

We basically adopt this last very open definition for our purpose of parsing a noun compound. This is mainly due to the fact that all of the possible constituents might actually appear in the construct as part of the 'core part' in a heading. It is obvious to see that more relaxed definitions of noun compounds would incur more difficulties in analyzing them. Since our main objective for this 'core part' analysis is to identify the 'core term', for most cases, we do not need to understand the semantic relationship between the constituents before picking out the 'core term' from them. That is what we have stated in our intention at the beginning of this subsection, that we need primarily a syntactic parsing of the noun compound without further resorting to complicated semantic analysis, if possible.

This simplified scheme saves us from investing large resources by applying various complicated techniques to understand the semantics in the compounds (Levi 1978, Leonard 1984, Warren 1978, McDonald 1982, Lauer 1995a, Lauer 1995b, Ryder 1994). These analyses and techniques require a large number of rules and/or corpus data to achieve moderate outcomes. Instead, we adopt a linguistic rule called "right-headedness" (Lauer 1995b:37), which can greatly simplify the process of identifying the 'core term' in a compound. Before introducing the rule, we will first describe a widely recognized linguistic pattern for a compound scenario.

In a typical noun compound, there are two roles to be played in the semantic representation: one is a noun that denotes the central subject or concept to be identified in that compound, while any other denotes a feature/characteristic of it, or a thing related to it. The former is called ***headnoun***, or *core term* in our previous expression. The latter is called ***modifier***. The ***modifier*** functions as a restriction or specialization so that the concept denoted by the whole compound is generally a subset of the concept denoted by the head only. For the example for ***Razor blade strip steel***, it denotes a subset of steel (a type of steel in terms of shape), which is then further restricted to a subset of it by its use (as 'razor blade'). A fine distinction can be made here. For many adjectives that act as modifiers, a formed compound still refers to the original head concept, except that some of its features (i.e. 'intensional attribute' in concept definition of chapter 2) have been specified. For instance, *Neutral spirits* is still spirits, with the modifier 'neutral' specifying the value of one of its attributes (the degree of alcohol).

The rule of 'right-headedness' specifies that in English the head of a compound is always the rightmost noun in it. This seems mostly true for our analysis of SIC headings and users' entries. Occasionally there is an exception. *Attorney general* is an example. Instances of these tend to be borrowed from other languages.

In the case when a gerund appears at the end of a compound, the whole compound would be treated as a process concept rather than an object concept. We then use the term 'predicate' to denote this gerund form. Example (2) illustrates this scenario.

(2) 3851 Ophthalmic lens grinding, except prescription--mfg
↑ ↑ ↑
modifier headnoun predicate

Noun Compound with Prepositional Phrase Attachment. Normally, a prepositional phrase adds complexity to a typical noun compound processing. However, the treatment for the SIC headings analysis can be rather simple as we have observed that the prepositional phrase part contributes less significant information to the overall 'core part' than the main compound does. Therefore, the prepositional phrase is treated as an amendment to the compound, i.e. its content can at first be ignored unless the main compound part needs more information for clarification, in which case the prepositional phrase part will then be parsed and treated as another modifier to the headnoun. Consider example (3):

(3) 3861 Flashlight apparatus for photographers, except bulbs--mfg

in the 'core part', the term *photographers* in the prepositional phrase can be treated as a modifier, if necessary. Then the whole 'core part' is equivalent to this: *Photographers flashlight apparatus.* The whole meaning is maintained after the transformation.

For our prototype implementation, the prepositional phrase is normally discarded unless there is a lack of modifier to the headnoun, in which case the terms in the prepositional phrase are treated as modifiers.

Compound Phrases with Conjunction. It is a more challenging task when there is more than one single phrase that constitute the 'core part'. The compound phrases we refer to here are those 'core parts' in which there is at least one conjunction (usually the word "and", or "or") that connects the individual phrases. In this part of compound phrases analysis, we assume a 'core part' has been pre-processed by the above-mentioned three types of scenario. In particular, the prepositional phrase part has been filtered out for this stage of consideration.

At first glance, it seems easy to solve this by just breaking apart the phrases at the point of the conjunction, and then treating each phrase as a separate and independent 'core part'. This happens to work for the example we gave for this category classification and those that only have three words left in the compound phrases after preprocessing (see example 4).

(4) a. 2296 Steel tire cord and tire cord fabrics--mfg
    b. 6211 Managers or agents for mutual funds

However, in many cases of the SIC headings, this simple treatment will not work. This is mainly due to the fact that in many occasions certain word(s) (either a modifier or the headnoun itself) is omitted in the description that might cause either incorrectly identifying the headnoun from the break-down phrases, or missing some modifier in the break-down. In either case, the information contained in the simple decomposed phrases is incomplete. This can be demonstrated in example (5), where the word in the bracket indicates the omitted term.

(5) a. 2515 Chair [springs] and couch springs, assembled--mfg
    b. 2675 Cardboard panels and [cardboard] cutouts--mfpm--mfg

For these types of scenarios, i.e. a four word (collocations are counted as one word) 'core part' containing a conjunction, we can still manage to parse syntactically the 'core part' by filling out the default given the position of the conjunction as illustrated in the above examples. This default-filling rule can be extended to apply to those 'core parts' that have the size of words more than four with the condition that the conjunction appears either at second from the beginning of the phrase or the second from the end, as in example (6). This would correctly generate two parallel phrases.

(6) 2675 Egg case fillers and [Egg case] flats, die-cut from purchased paper-- mfg

Another view of this rule is that the two words surrounding the conjunction are treated as if they were one word so that the original compound phrases can be processed as a single phrase. We will generalize this view in a moment.

However, occasionally there are some exceptions to this. Example (7) will fail our default-filling rule. In this case, more in-depth (i.e. semantic) analysis is needed to resolve the problem. The phrase after the conjunction indicates a much more general concept, hence the concept before the conjunction should be repeated after that.

(7) 5072 Locks and [lock] related materials--wholesale

When the position of a conjunction does not fall in the precise region required by the default-filling rule, we need a deeper analysis of the context of the conjunction. We now introduce a 'bracketing' rule. It is aimed at 'bracketing' a certain context of the conjunction so that this bracketed region will function as if it were one single word/concept in the overall 'core part'. The premise for the 'bracketing' rule is that there exist parallel concepts in that contextual region. The criteria for the resultant bracketed region are as follows:

- *Parallelism.* Both parts should be similar in content, and perhaps symmetrical in representation (morpheme patterns), if possible.
- *Minimalism.* The resultant concepts are kept as 'small' (i.e. the shortest length of parallel concepts) as possible.

For the specific implementation, we first apply grammar to find the symmetric parallel patterns, then verify them by semantic checking. A sketch of the algorithm is as follows:

1. Start from the center of the conjunction in a whole part;
2. Expand the bracket region to both sides one word/term at a time;
3. Check with the parallel criterion, and repeat step 2 if not satisfied.

Examples in (8) show how gerunds help determine the boundary of the bracketed region. The square bracket indicates the final bracketed area.

(8) a. 3567 [Core baking and mold drying] ovens--mfg
b. 3567 Paint [baking and drying] ovens--mfg

Example (9) needs a more complete semantic examination to determine the bracketing boundary.

(9) 3694 Ignition [cable sets or wire assemblies] for internal combustion engines--mfg

There is a more complex scenario in the compound phrase analysis when more than two parallel concepts exist in a subcomponent. In example (10), there are three parallel concepts functioning as headnouns. To generate the correct subcomponents in the parsing, we need to keep track of all the punctuation marks. In this case, when we see the conjunction *and* after a comma which signals multiple parallel concepts, we then need to backtrack the term *chloride* that has been categorized as a 'descriptor'. Now all three terms, including the term *hypochlorite* after the conjunction *and*, will be correctly categorized into the 'core part'.

(10) 2819 Calcium carbide, chloride, and hypochlorite—mfg

Notice that this rule will be able to tell the difference between potential multiple parallel concepts, and two really separated subcomponents, as in example (11). In the latter case, there is no comma before the conjunction *and*.

(11) 2091 Chowders, fish and seafood: canned--mfg

In general, the techniques we present here are the type of weak techniques that do not aim at generating perfect analysis results. They are simple and efficient, and can produce results that are satisfactory in most cases. Table 6.2 presents a sample of some intermediate results in the analysis of candidate SIC headings for the query ***Aluminum Containers***. It can be observed that most of the content-bearing words are correctly classified. Occasionally, there are some errors. For example, 'including' is wrongly recognized as a predicate in the third example.

SIC heading	Core Term	Core Modifier	Predicate	Descriptor	Constraint
3334 Extrusion ingot, aluminum: primary—mfg	ingot	extrusion		aluminum	primary
3365 Aluminum and aluminum-base alloy castings, except die castings—mfg	castings	aluminum, aluminum-base, alloy		die castings	
3363 Aluminum diecasting, including alloys—mfg	aluminum		diecasting , including	alloys	
3471 Coloring and finishing of aluminum and formed products, for the trade—mfg	products	aluminum, formed	coloring, finishing		
6051 Aluminum bars, rods, ingots, sheets, pipes, plates, etc.--wholesale	bars, rods, ingots, sheets, pipes, plates,	aluminum			

Table 6.2 Some sample results of SIC headings analysis

6.3.2 SIC headings Retrieval

The second major step of business category classification/search subsystem is to fetch and rank SIC headings to respond to a user's query. The input is the parsed user's entry as well as the SIC headings generated in the first step. The output is a ranked list of potential relevant SIC headings. An algorithm is constructed to find industry headings to best match the user's entry. The essence of it is to make similarity comparisons only between the corresponding decomposed parts of the query and SIC headings that are functionally equivalent. In this algorithm, we use a general lexical taxonomy (WordNet) to aid the semantic expansion required in this matching process. We denote the complete set of the SIC headings as H_0, and the final resultant set of headings as H_R , where $H_R = \phi$ initially. t_{ijk}^{q} refers to the *k*th term in the subcomponent *j* of field *i* (i= industry/business category) in a query.

Step 1. Generate a list of candidate terms { t_{ijk}^{q} } from the result of query parsing. The list is sorted by the 'importance' contribution[11] of each term to the query.

Step 2. Locate all the headings from H_0 to produce a candidate heading set H_I that contains this next term t_{ijk}^{q} (and its synonyms) from the term list. If $H_I = \phi$, go to Step 4.

Step 3. For each heading in H_I, estimate the similarities of all the decomposed parts (e.g. headnoun, modifier, descriptor, and constraint) between the heading and the corresponding counterparts in the query. Compute the overall similarity value between the two. Merge H_I into H_R, and rank H_R by the descending order of the overall similarity value.

[11] A term's information content value can represent this importance contribution.

Step 4. If $H_I = \phi$ or the result set H_R is not satisfying (e.g. the top-ranking similarity values are below a threshold), replace t^q_{ijk} with its morphological variants (e.g. via a pluralization rule), and/or with its semantic variants (e.g. the superordinate concept in the IS-A hierarchy), and repeat Step 2.

Step5. If the term list $\{ t^q_{ijk} \}$ is not empty, go to the next one and repeat Step 2. Otherwise, output the resultant set H_R.

The above procedure describes a conceptual approach to matching an industry class with the help of both lexical and semantic knowledge. It first uses grammatical rules to analyze the query terms and break them down into functional units, then it performs conceptual matching for each functional unit against the counterpart from a potentially relevant heading. The logical structure of a taxonomy is used for exploring the conceptual connections. Through tracing the relationships among concepts/classes this algorithm is able to draw out a user's implicit yet conceptually relevant query term(s) based on the explicit term concept it contains.

In Step 3 of estimating the similarity value for each component part, there is a priority in performing this task. Basically, the headnoun is first compared and the candidate heading will be discarded if the value for the headnoun similarity is below a certain threshold. This design philosophy came from two perspectives. First, we see the inherited comparison in retrieval as a pair of objects/concepts comparison, that a query or heading is treated as one though mostly a complex object/concept with many features/attributes, and the headnoun is the centerpiece of such an object/concept. It is mainly the headnoun that designates this complex object/concept in a semantic space, where other parts play roles as intensional attributes of such object that 'fine-tune' the object's position. Secondly, the structural representation of the SIC headings also stresses the role of the headnoun, as it is the only part that cannot be omitted in generating the parsing result. The premise of such a design is the correct identification of the headnoun, which is of course not guaranteed by our 'shallow' heading analysis rules in the last subsection. Therefore, it runs the risk of generating irrelevant parts for comparison; hence the overall performance would be affected. A practical outcome of this strategy is that it will improve the retrieval efficiency by filtering out deemed non-relevant candidates at an early stage.

It should be pointed out that this is yet a rather coarse outline of the solution procedure. Implementation of specific steps will be addressed in the coming discussion. Some additional steps may also be added to improve the overall performance. The experiments we conduct in section 6.4 will consider various factors to encompass more complete and fine-tuned scenarios. At this moment, we will discuss some of the important details in the algorithm. Specifically they are: a) the determination of local subcomponent similarity and, b) the calculation of global overall similarity.

6.3.2.1 Determination of Local Subcomponent Similarity

Determination of subcomponent similarity can be generalized as the determination of phrase-phrase similarity since the decomposed subcomponent part tends to be a simple phrase like structure.

For comparison within a subcomponent, it would be desirable to compare those terms that are functionally or semantically correspondent in a subcomponent. This would be particularly useful for the modifier(s) comparison. However, since we did not pursue further semantic analysis for a decomposed part (particularly the noun compound or phrase in the 'core part'), where this semantic analysis alone can be complicated enough to render a separate study, this leaves us to resort to the simple, iterative terms comparison approach. In this regard, there are two main tactics to perform this iterative comparison procedure.

The first is a simple and brute-force approach that regards the subcomponent as a list of concatenated unstructured terms ("a bag of words"). It adds all the possible pairwise similarities between a term in an entry and a term in the heading for that specific subcomponent, normalized by dividing by the number of terms in the corresponding subcomponent of both the entry and the heading, i.e.

$$Sim(t^q_{ij\cdot}, t^c_{ij\cdot}) = \frac{1}{n^q_{ij\cdot} * n^c_{ij\cdot}} \sum_{k=1}^{n^q_{ij\cdot}} \sum_{l=1}^{n^c_{ij\cdot}} Sim(t^q_{ijk}, t^c_{ijl})$$

A somewhat closer to the desired scenario is an approach that assumes that there is a structure (although unknown) in each part of the entry/heading. Therefore for each term in the entry part, there exists a corresponding functionally equivalent term in the heading's part. Since we do not know where exactly the corresponding term is, we can arguably determine it by picking the term that has the highest similarity value with the term being matched. The formula is as follows:

$$Sim(t^q_{ij\cdot}, t^c_{ij\cdot}) = \frac{1}{n^q_{ij\cdot}} \sum_{k=1}^{n^q_{ij\cdot}} \max_l [Sim(t^q_{ijk}, t^c_{ijl})]$$

Choosing the best similarity pair augments the use of formula 4.3, where word pair similarity is determined as the best sense pair similarity among all the possible pairwise sense similarities.

6.3.2.2 Calculate the Overall Similarity for the Candidate Heading

In the spirit of Gower's general function of global weighted feature similarity (formula 2.17), we can represent the overall similarity method for calculating the user-entered business category and a SIC heading candidate as the weighted average of each subcomponent similarity:

$$Sim(F^c_i, F^q_i) = \frac{\sum w_{ij} Sim(t^c_{ij}, t^q_{ij})}{\sum w_{ij}},$$

where subscript i at this time denotes a business category field, j refers to the subcomponent part from the entry and ranges over the decomposed subcomponent part(s) for each specific query; w_{ij} is the weight for each decomposed part j comparison. Weighting can be fixed for each subcomponent part for all entries, or can be dynamically determined based on the relationship among the subcomponents for a specific entry.

The dynamic weight assignment for each subcomponent part varies according to the actual degree of importance each part has in a particular entry. The degree of importance for a part can be viewed as the amount of information it contributes to the whole entry, which is measured by the value of its information content. Thus it can be argued that a component part with high information content should be weighted more in an entry. Therefore the weight for a component part is the average of the information content values for all the terms in that part.

$$w_{ij}^{q} = \frac{1}{n_{ij}^{q}} \sum_{k} IC(t_{ijk}^{q})$$

where k ranges over the number of terms (n_{ij}^{q}) in a subcomponent j.

6.4 Experimental Design

We now have a complete implementation of the function for a phrase-like business catalog retrieval—a subsystem defined in the overall EID design. The remainder of the chapter describes a range of experimental work aimed at evaluating both the category parsing strategy and the overall similarity determination method. Classical IR performance measures are used to evaluate the model.

In this section, several requirements for the experimental setup are discussed.

6.4.1 The Lexical Taxonomy

As described in section 6.3, the final decomposed concept comparison for the task of subcomponent similarity computation requires calculating the semantic similarity between a single elemental pair of concepts. The use of the WordNet as a lexical taxonomy tool would be appropriate for this task. Our previous concern was that since the genre of SIC headings is primarily business, a comprehensive trade-oriented taxonomy would be ideal. Unfortunately we have not found any (nor even close) such taxonomy. The experience of using WordNet for SIC headings search (as will also be discussed later in 6.6.2) has also demonstrated its value as applied to this particular domain.

6.4.2 The Corpus

Since we are using the WordNet as the taxonomy to aid retrieval in similarity analysis, the SemCor corpus that is annotated based on the WordNet sense definitions remains a natural candidate. As we discuss earlier in Chapter 4, the main drawback of SemCor is its relatively small size. To deal with the problem of sparse data, we again employed the Good-Turing method to adjust the frequency estimate of the SemCor corpus. As discussed in chapter 4, the resultant data performed similarly to others that use much larger corpora.

6.4.3 The SIC Headings

As determined in section A.3.2 , we use the U.S. government 1987 version SIC codes for the implementation of the directory indexing structure. As well, to serve the purpose of having SIC codes as comprehensive as possible, we include all the subheadings under the four-digit industry class headings. This generated a total indexing size of 17,802 headings and subheadings.

The electronic source of the SIC codes was obtained from an FTP site at a company called Vancouverweb[12]. Since the original source was OCR scanned, there were many errors in the electronic version, most of which were corrected by verifying manually with the printed official SIC manual. Also, we retained the original style of this electronic copy, where an abbreviation of the industrial type information is usually attached at the end of each heading, for example, *3316 Staples, steel: wire or cut—mfg.*

[12] http://vancouver-webpages.com/

6.4.4 The Test Data

To reflect the scenario of the real world application, the test data were selected from an international trade mailing list called Trade-L. Trade-L is a daily list published by Tradewinds Publishing Company in the U.S.[13]. Each email of Trade-L consists of a list of user-advertised trade leads that are categorized into one of three main categories: Offer, Demand, and Miscellaneous. Each piece of trade lead information contains a structure of several fields such as product/service name, product/service description, business type (assigned by the advertiser, e.g. offshore firm, exporter), contact information (person name, address, email, web site, etc.), opportunity type (a total of 40 types designed by the system, e.g. *T105=Offer-manufacturing, T209=Demand-capital investment*), target country, etc. In general, the structure of the trade advertisement is similar to the record structure of our EID database design. In particular, the brief product/service name field can be treated as the counterpart of SIC headings in the EID database.

We randomly selected about 150 trade lead postings as user query input for our test, avoiding postings that we had examined earlier to design our retrieval algorithm. Since the original Trade-L postings were not designed with the purpose to direct automatic retrieval of trade leads, the quality of the content representation is low in terms of the requirement for an automatic process. Also Trade-L is intended to function as a medium to provide freely available trade information with the least manual interference, therefore the noise in the posting data tends to be very high. To ensure the consistency of the format of the query (test) data with the counterpart of SIC headings upon which our analysis and design are based, we have inspected each randomly selected posting to filter out obviously inappropriate data parts or the complete posting. Some of the concerns are as follows:

- Discard minimal content bearing words, e.g. the word *export* in "Auto Parts Export", and *quality* in "Quality digital watches".

[13] http://www.intl-trade.com

- Remove product specification in numerical quantity, e.g. the purity specification in "Cadmium 99.999% Purity".
- Delete entries that specify a specific brand product, or a specific geographical region, e.g. "Compaq Deskpro 2000", "555 & Marlboro", "Cars from Korea".
- Drop entries with jargon or abbreviations, e.g. "DAP", "SSP", "Quantum Atlas II XP32275W".
- Discard entries that contain more than one unrelated topic in one basket, e.g. "Steam Coal - Plywood - Rattan".
- Drop entries that are too obscure to understand even for a human subject without further reading the product description part, e.g. "Naptha", "Engerix".
- Delete repeated entries.

Some of the above mentioned entries can actually be dealt with given more powerful rules in the parser. Since this is beyond the focus of this SIC headings search experiment, we do not include them for testing. After the above filtering, we have 95 valid user entries as the testing data. A complete list of all the valid entries can be found in Appendix A.

For each of these 95 test entries, we manually determined all the relevant heading(s) to form the answer set. This procedure is time-consuming but less than the tasks in typical IR work, since the relevant answers are limited to potential categories in the whole classification hierarchy. Logically, there should be only one most suitable heading under which an entry should be linked. In reality, there are often cases where several headings have exactly the same or very similar content in terms of their descriptions, only different in their presentations (typically switching the headnoun modifier with the descriptor); see example (12).

(12) 2842 Degreasing solvent--mfg
2842 Solvents, degreasing--mfg

There are also cases where everything in the main subcomponents are the same except the industry type part, which is typically the case when the industry type is not specified or requested in a user's entry. See example (13):

(13) 5992 Flowers, fresh--retail
     5193 Flowers, fresh--wholesale

In addition, the users' entries tend to be short compared with the rather rigorous and specific specification in the SIC headings (see Table 6.3). This can often result in a partial matching. We relax the criteria by considering the default as a relevant answer as long as the whole concept does not deviate significantly. As in example (14), we consider both as relevant headings to the query *Organic Chemicals.*

(14) 2869 Laboratory chemicals, organic—mfg
     2869 Organic chemicals, acyclic--mfg

Hence the strategy we adopt here transforms the task from a purely classification process to a more retrieval like job which allows us to adopt the standard evaluation mechanism used in a typical IR job, which is to calculate precision values for each standard recall point.

Statistics for the collections of headings and query entries are given in Table 6.3.

Measure	Result
Total number of SIC headings	17,802
Average number of words per heading	4.88
Total number of usable query entries	95
Average number of words per query	2.57
Average number of relevant headings per usable query	4.19

Table 6.3 Some Statistics for both SIC headings and query collections

6.4.5 Term Similarity Measure

From the discussions in the previous chapters, there are many alternatives to choose from when a specific word-level similarity determination method is required. Since the focus of this experiment is on the evaluation of higher-level semantic retrieval, we need to select one that fits well in the overall similarity formula. For example, those having a [0,1] value range would be a perfect fit for the proposed subcomponent and overall similarity formulae in that the resultant overall similarity score is also in this range. Both the Dice and the Jaccard measures have this feature. We select the Dice measure since it has shown better performance in the previous experiments.

6.4.6 Evaluation Measures

The major evaluation technique we adopt here is the standard methodology used in the Text Retrieval Conference (TREC) for document retrieval (Harman 1995). This is the so-called 11-point recall-precision graph. To measure retrieval effectiveness, precision values are calculated at each cutoff recall point for the result of a given query, they are then interpolated to a standard 11 recall points and averaged over all the query results.

Before calculating the interpolation, for every query q, there is a need to perform an adjustment of the precision value $\mathrm{Pre}(q)$ to a new estimate $\overset{\wedge}{\mathrm{Pre}_q}(\rho)$ for each recall value $\rho \in [0,1]$, which is given in equation 6.1 (Schauble 1997:24):

$$\overset{\wedge}{\mathrm{Pre}_q}(\rho) := \max\{\mathrm{Pre}(q) | \mathrm{Rec}(q) \geq \rho\} \qquad (6.1)$$

This adjustment is meant to define the precision at a recall level ρ as the maximum precision value achieved for all the recall values $\mathrm{Rec}(q)$ that are no less than ρ. This will ensure a monotonically decreasing precision-recall curve where each recall value corresponds to a unique precision value.

In addition to the standard 11-point recall-precision graph measure, we also include some other measures that are frequently used in assessing an IR task. They are described as follows:

P_{avg}: a single precision value which is an average precision value over all the retrieved relevant headings;

$P_{3\text{-point}}$: a three-point average precision which is the average of precision values at three typical recall points: 0.2, 0.5 and 0.8. Each of the three points represents typically a precision-oriented, an intermediate, and a recall-oriented performance preference, respectively;

$P_{(r)}$: an original (non-adjusted) precision value calculated at the *r*th (*r*=5, 10, 20, 30) result heading among the ranked resultant set;

R_{avg}: an average recall value of all the queries. This is achievable since we are able to provide the relevant set for each query.

6.5 Results

The benchmark for our model is the classic statistically based vector space model using the inverted document frequency (tf*IDF) method. This simple yet proven technique has been widely used in various domains of application and is still the basis upon which many more sophisticated models are built. It was also employed as a benchmark comparison for matching extremely short documents, such as image caption retrieval (Smeaton and Quigley 1996). In determining tf*IDF, the weight of a term in a query is calculated the same way as that in a document/heading, since both are short and have relatively the same scale. The popular cosine coefficient is them used to calculate the overall similarity between them.

To ensure a fair level of comparison between the vector space model and our semantic similarity model, we allow words' morphological variants to be considered as alternatives in the vector space model. This has an effect similar to the stemming procedure applied directly to the original text (i.e. headings). In addition, typical stop words and particular functional words in the SIC headings (e.g. those industry type indicators at the end of most headings) are filtered out to reduce the noise in the calculation.

For each of the 95 queries, a ranked list of results were generated according to each matching scheme. The data from the top 30 results are used to calculate the various performance measures we defined in the last section.

The corresponding performance data for the vector space method (Run1) and the semantic similarity method (a specific test Run3) are listed in Table 6.4. In general, it can be observed that for each performance criterion, there is a certain degree of improvement for the proposed semantic similarity model over the classic vector space model. The average precision has increased from 55.2% to 65.0%. The average recall has increased from 45.3% to 55.8%. Notice that the interpolated 11-point precision values are significantly higher than the rank point precision values, which are directly calculated without any adjustment. The reason for this is that the answer set (the relevant headings for each query) is relatively small[14] (average 4.19 per query, see Table 6.3). Hence the adjusted precision values tend to be 'bumped up' for queries when the number of relevant results in the resultant set is small (say one or two).

To see whether the observed differences in the recall-precision graph are really meaningful or simply due to chance, we need to conduct a statistical test. In this context, we wish to compare the proposed similarity-based approach (Run3) with the baseline vector space model (Run1). Since the performance differences are much greater between queries than between methods, measurements must be viewed as matched pairs for analysis (Hull 1993). The most common approach to this is to apply the paired *t*-test to these differences. With *n*=95, the resultant *t*-score is 2.23, which indicates the observed difference of precision values between the proposed model and that of the vector space model is significant at 0.025 level.

[14] This is understandable given the nature of the task as SIC retrieval. Ideally, there should be only one correct answer where the query fits into such a category. In most cases there is often more than one resultant heading that is semantically related to the query.

The improvement of recall performance is also in line with our expectation. It should benefit from the semantic expansion in our retrieval algorithm.

	Run1	Run3
P at 0.00	0.688	0.792
0.10	0.688	0.792
0.20	0.664	0.765
0.30	0.656	0.757
0.40	0.614	0.719
0.50	0.529	0.623
0.60	0.520	0.613
0.70	0.474	0.562
0.80	0.457	0.544
0.90	0.427	0.536
1.00	0.427	0.535
P_{avg}	0.552	0.650
$P_{3\text{-point}}$	0.550	0.644
R_{avg}	0.453	0.558
$P_{(5)}$	0.317	0.385
$P_{(10)}$	0.208	0.251
$P_{(20)}$	0.134	0.154
$P_{(30)}$	0.100	0.110

Table 6.4 Performance values for Run1 (vector space) and Run3 (semantic similarity)

A clearer observation of the performance difference can be viewed from the recall-precision graph in Figure 6.2.

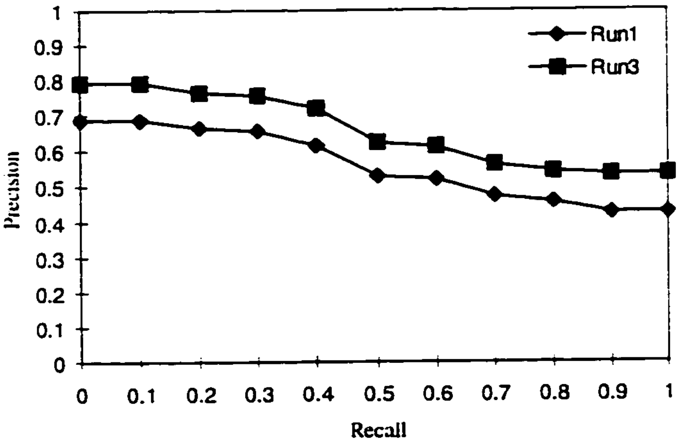


Figure 6.2 Recall-precision graph for Run1 and Run3

6.6 Comparisons

In section 6.4 where we discussed the design implementation of the SIC headings retrieval, there are several decision factors that need to be addressed. In this section we run experiments to compare various alternatives for the specific implementation regarding the decision factors for each of the following three dimensions:

1. Local within subcomponent similarity determination
2. Global between overall similarity determination
3. Parsing vs. non-parsing model

6.6.1 Local Within-Subcomponent Similarity Determination

The subcomponent phrase-like similarity comparison is essentially a process of many-to-many single-term pairwise comparisons. One common approach is to treat it as a series of one-to-many comparisons. The latter is then further treated as a series of one-to-one pair comparisons. For each of the decomposed comparison steps, there are basically two comparison methodologies. First is the 'average of all' approach, where the value of the current step is calculated as the average value of all the results among the decomposed step. The second approach is the 'best of all' approach, where the current step value is equal to the best value among all the results of the decomposed step. Here again it is an augmentation to formula 4.3 where word pair similarity scheme is similarly deployed. When applied to our query and headings matching, the combination of these two generates four possible scenarios listed in Table 6.5.

	Query	Heading
Run2	Average	Average
Run3	Average	Best
Run4	Best	Average
Run5	Best	Best

Table 6.5 Combination of schemes for subcomponent similarity calculation

The specific determinations for each of the four runs correspond to the following equations:

$$Run2\_Sim(t^q_{ij\cdot}, t^c_{ij\cdot}) = \frac{1}{n^q_{ij\cdot} * n^c_{ij\cdot}} \sum_{k=1}^{n^q_{ij\cdot}} \sum_{l=1}^{n^c_{ij\cdot}} Sim(t^q_{ijk}, t^c_{ijl})$$

$$Run3\_Sim(t^q_{ij\cdot}, t^c_{ij\cdot}) = \frac{1}{n^q_{ij\cdot}} \sum_{k=1}^{n^q_{ij\cdot}} \max_{l}[Sim(t^q_{ijk}, t^c_{ijl})]$$

$$Run4\_Sim(t^q_{ij\cdot}, t^c_{ij\cdot}) = \frac{1}{n^c_{ij\cdot}} \sum_{l=1}^{n^c_{ij\cdot}} \max_{k}[Sim(t^q_{ijk}, t^c_{ijl})]$$

$$Run5\_Sim(t^q_{ij\cdot}, t^c_{ij\cdot}) = \max_{k,l}[Sim(t^q_{ijk}, t^c_{ijl})]$$

The results of these four combinations are listed in Table 6.6. The results of Run2 with the calculation of the averages of all pairwise comparisons show the weakest performance values among the four combinations. This is not surprising as values from non-corresponding term comparisons may generate noise for the real corresponding part comparisons as the result of the averaging effect. For instance, example (15) samples two high ranking resultant headings (each prefixed with its similarity score) for the query *Golf Balls*. Intuitively, the second heading should be a more relevant result than the first one. However, due to the averaging effect resulting from the core modifier similarity calculation, the overall similarity score of the second result ends up lower than the first one.

(15) a. 0.862 3949 Bags, golf--mfg

b. 0.830 3949 Balls: baseball, basketball, football, golf, tennis, pool, and bowling--mfg

The advantage of this 'average of all' approach is that when there is a tendency of mismatching, it will reduce such a likelihood by averaging all the pairwise comparison values. In the example (16) query *Cashew Nuts*, the headnoun *nuts* in the query has some significantly different senses. When determining the similarity for the modifier part, the average method of comparing the word *cashew* with each term in the heading would result in a reduced similarity value than that in the 'best of all' matching approach.

(16) 5072 Bolts, nuts, rivets, and screws--wholesale

	Run1	Run2	Run3	Run4	Run5	Run6	Run7
P at 0.00	0.688	0.750	0.792	0.766	0.778	0.791	0.787
0.10	0.688	0.750	0.792	0.766	0.778	0.791	0.787
0.20	0.664	0.741	0.765	0.756	0.752	0.784	0.780
0.30	0.656	0.718	0.757	0.729	0.738	0.757	0.756
0.40	0.614	0.661	0.719	0.675	0.706	0.692	0.690
0.50	0.529	0.581	0.623	0.600	0.616	0.600	0.592
0.60	0.520	0.573	0.613	0.593	0.606	0.593	0.585
0.70	0.474	0.535	0.562	0.549	0.549	0.552	0.541
0.80	0.457	0.525	0.544	0.539	0.526	0.536	0.528
0.90	0.427	0.511	0.536	0.526	0.511	0.512	0.504
1.00	0.427	0.511	0.535	0.526	0.510	0.512	0.504
P_{avg}	0.552	0.611	0.650	0.625	0.633	0.636	0.631
$P_{3\text{-point}}$	0.550	0.615	0.644	0.632	0.632	0.640	0.633
R_{avg}	0.453	0.516	0.558	0.526	0.579	0.547	0.547
$P_{(5)}$	0.317	0.361	0.385	0.366	0.377	0.385	0.377
$P_{(10)}$	0.208	0.226	0.251	0.231	0.252	0.247	0.236
$P_{(20)}$	0.134	0.146	0.154	0.148	0.157	0.153	0.150
$P_{(30)}$	0.100	0.106	0.110	0.107	0.113	0.111	0.110

Table 6.6 Results of various subcomponent similarity determination methods

Compared with SIC headings, the length of the user's entry is relatively short (see Table 6.3), and hence less noise will occur when using each term in the query to lead to the one to many matching in the heading, than the reverse scenario of using the term in the headings to lead to the one to many matching in the query. Therefore we observe Run3 has a better performance than Run4.

Run5 is tested just for the sake of completeness in the enumeration of all the possible combinations. The logic behind it appears less appealing. Nevertheless, the results are not much worse than the best-average (Run3) method. It even has the best average recall value among all. This is mainly due to the fact that the lengths of both queries and headings are short, and they are even shorter when both are decomposed into a subcomponent part in order to compare. Therefore, when a particular salient factor appears in a heading, it will dominate the contributions to the overall similarity score.

In re-examining the nature of the formulas, we can observe that formulas for Run2 and Run5 use symmetric measures for the within-subcomponent similarity calculation — i.e. $Sim(t^q_{ij.}, t^c_{ij.}) = Sim(t^c_{ij.}, t^q_{ij.})$, whereas the formulas for Run3 and Run4 use asymmetric measures as they sum over all the terms in only one of the subcomponents (either in the query or the heading). In fact, formulas for Run3 and Run4 themselves are symmetrical — i.e. $Run3\_Sim(t^q_{ij.}, t^c_{ij.}) = Run4\_Sim(t^c_{ij.}, t^q_{ij.})$. The fact that Run3 performs better than the rest provides some support to the arguments made by others that asymmetric similarity measures are often appropriate (e.g. formula 2.8).

Some more runs were conducted with the idea of fusing and combining some of the above four basic runs. To form a symmetric matrix of many to many comparisons, Run6 fuses Run4 and Run5 simply by adding the scores of each calculation. This is done in the belief that fusing may take advantage of each calculation from the asymmetric matrix data (Smeaton and Quigly 1996). The results, however, don't seem to reveal any expected improvement at this subcomponent level comparison.

$$Run6\_Sim(t^q_{ij.}, t^c_{ij.}) = Run3\_Sim(t^q_{ij.}, t^c_{ij.}) + Run4\_Sim(t^q_{ij.}, t^c_{ij.})$$

Similarly, after observing some sample results, we found that the resultant set from the average-all approach (Run2) is often somewhat different from that of the average-best approach (Run3). We decided to combine them together in Run7. Again, the results are not supportive of this fusion idea.

$$Run7\_Sim(t^q_{ij.},t^c_{ij.}) = Run2\_Sim(t^q_{ij.},t^c_{ij.}) + Run3\_Sim(t^q_{ij.},t^c_{ij.})$$

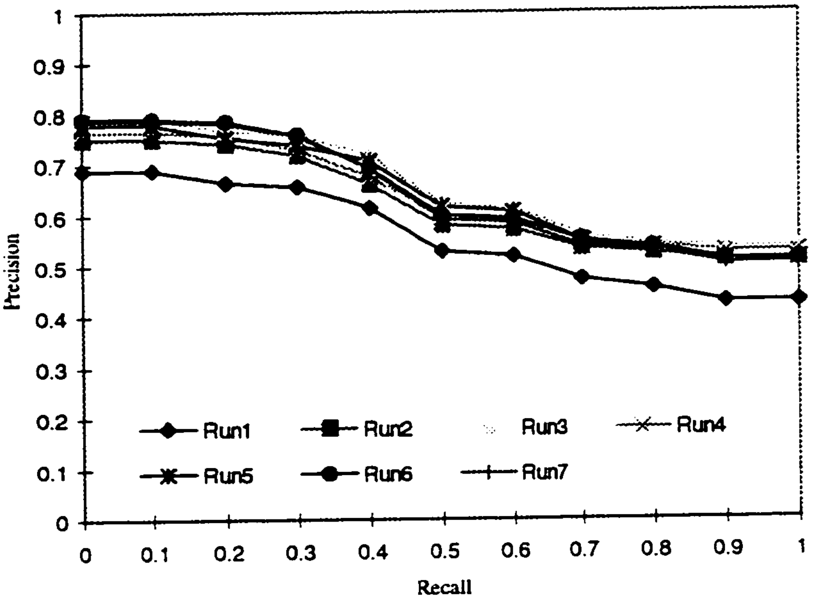


Figure 6.3 Recall-precision graph for subcomponent similarity comparison methods

In general, for the subcomponent similarity determination, the best performance is achieved with the 'average-best' approach (Run3). Attempts to fuse some of the basic approaches do not provide any better results. The results of all these methods are drawn in Figure 6.3. For comparison, the results of the benchmark vector space model (Run1) are also included. It can seen that there are varying degrees of improvement over the benchmark.

Once again, we need to conduct a statistical analysis to compare the precision values among these 7 approaches. Similar to the comparison for two means in Table 6.5, we use the two-way ANOVA of randomized blocks design to compare these seven means. The block effect accounts for the variances among queries.

Source of Variation	SS	df	MS	F	P-value	F critical
Queries	43.92922	94	0.467332	13.00227	6.61E-93	1.279224
Runs	0.589317	6	0.09822	2.732696	0.012623	2.114639
Error	20.27149	564	0.035942			
Total	64.79003	664				

Table 6.7 Two-way ANOVA for the runs in Table 6.6 at 5% level

The corresponding ANOVA is presented in Table 6.7. Note that the F-ratio for the groups of runs is 2.73 with 6 and 564 df and this is significant beyond the 0.05 level, clearly indicating differences in the precision values of these 7 runs.

Since the analysis indicates a significant difference in the Run means, we would be interested in multiple comparisons to discover which Run's means differ. We use Duncan's multiple range test to perform this test (Montgomery 1991). The standard error of a treatment (Run) mean for the number of blocks b is

$$S_{\overline{Run_i}} = \sqrt{\frac{MS_E}{b}} = \sqrt{\frac{0.035942}{95}} = 0.019451$$

From Duncan's table of significant ranges, we can obtain the values $r_{.05}(p, df_E)$ for range step p=2, 3, ... , 7, and the number of degree of freedom for error df_E=564. Convert these ranges into 6 least significant ranges:

$$R_p = r_{.05}(p, df_E) S_{\overline{Run_i}} \quad \text{for } p = 2,3,\ldots 7$$

Thus, the least significant ranges are

$$R_2 = r_{.05}(2{,}564) S_{\overline{Run_i}} = 2.77 * 0.0195 = 0.0539, \ldots, R_7 = 0.0620$$

By arranging all the Run means in ascending order, we can access the significance of observed differences between them. This is done by a sequence of pairwise comparisons of mean difference against a certain least significant range R_p. We begin with the largest (Run3) versus the smallest (Run1) by comparing their difference against R_7 as follows:

$$\text{Run3 vs. Run1:}\quad 0.650 - 0.552 = 0.098 > 0.062(R_7)$$

which indicates the difference is significant. Next, the difference of the largest (Run3) and the second smallest (Run2) is computed and compared with R_6. The process is continued until the differences of all possible 21 pairs of means have been considered.

After all the pairwise comparisons of the difference of the run means, we obtain the conclusions that Run1 is significantly different from all other six runs. And there are no significant differences among these other six runs.

6.6.2 Global Between-Subcomponent Similarity (Weight) Determination

As we have discussed previously, there are basically two schemes in determining the weight of a subcomponent similarity value in the overall similarity calculation: static and dynamic. In the static weighting scheme, there is a fixed weight value for each subcomponent part, and it will remain the same for each incoming user's entry no matter whether a particular subcomponent part exists or not. For dynamic weighting, the weight factor depends on the resultant subcomponent parts generated from the parsing algorithm, hence it varies from query to query.

There are two categories that can be considered for the dynamic weighting scheme: equal weights and variant weights. The first scheme assigns an equal weight factor to each subcomponent, the value of which is equal to the inverse value of the number of subcomponents generated from parsing a specific query. Notice that this is different from the static weight assignment method as the actual weight value for a part will not be determined until the parsed subcomponents are generated. In fact, this equal weight scheme is essentially a no-weight scheme regarding the objective of arriving at an overall similarity value.

A varying weight assignment seems more logical as the importance of a part relative to others in the entry should be adequately reflected. To determine this relative 'importance' for a particular part in a query, we decided to use the average information content of a subcomponent to represent this relative weight determination. The aim is to provide greater weight to a part that conveys more specific information.

As for the determination of the information content, there are a further two methods. The first is a domain-independent approach where the frequency information comes from the mixed genre corpus SemCor. This would come in handy as we already use the SemCor corpus to determine the information content value for a specific synset in WordNet. The second is the domain-dependent approach where the frequency data come directly from the SIC headings statistics.

A diagram of the classification of these weighting schemes is indicated in Figure 6.4.

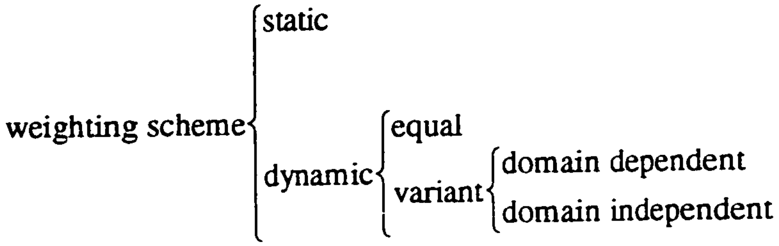


Figure 6.4 Subcomponent weighting schemes

Run8 and Run9 test two typical static weighting schemes. Since the headnoun is regarded as an important indicator in a business category description, our tests of static weight assignment will center on the headnoun weighting. Run9 is a pro-headnoun scheme where 50% of the weight is given to the headnoun and the other 50% to all other three subcomponent parts. In Run8, the scenario is reverse, where only 20% weight is given to the head noun, and 80% to others. The results from Run8 and Run9 in Table 6.8 confirm our conjecture, that the headnoun does play a central role in determining the overall heading's relevancy.

In the dynamic weighting scheme, Run10 tests the simple equal weighting approach. The weight factor is defined below:

$$Run10\_w_{ij}^{q} = \frac{1}{n_{i.}^{q}}$$

For the varying weighting scheme, Run11 uses the SIC data as the source to determine the information content value. Notice that the weight factor in this definition needs to be normalized.

$$Run11\_w_{ij}^{q} = \frac{1}{n_{ij}^{q}} \sum_{k} IC(t_{ijk}^{q})$$

Run3 is repeated here to represent using the SemCor corpus to calculate the subcomponent information content value. The actual information content value $IC(t_{ijk}^{q})$ for the term t_{ijk}^{q} is calculated as the weighted average of all its senses' information content values. This inner layer weight factor is based on the sense frequency information.

For each of the runs in the inter-part weight determination, the intra-part (i.e. subcomponent) similarity calculation is based on the best performance method (Run3) obtained from section 6.6.1.

	Run8	Run9	Run10	Run11	Run3
P at 0.00	0.743	0.766	0.784	0.760	0.792
0.10	0.743	0.766	0.784	0.760	0.792
0.20	0.718	0.760	0.760	0.735	0.765
0.30	0.710	0.752	0.752	0.721	0.757
0.40	0.692	0.708	0.716	0.697	0.719
0.50	0.618	0.613	0.622	0.619	0.623
0.60	0.614	0.610	0.611	0.604	0.613
0.70	0.568	0.572	0.557	0.553	0.562
0.80	0.543	0.557	0.539	0.531	0.544
0.90	0.532	0.549	0.531	0.522	0.536
1.00	0.531	0.549	0.531	0.522	0.535
P_{avg}	0.625	0.642	0.644	0.627	0.650
$P_{3\text{-point}}$	0.626	0.643	0.640	0.628	0.644
R_{avg}	0.484	0.526	0.558	0.547	0.558
$P_{(5)}$	0.362	0.378	0.383	0.385	0.385
$P_{(10)}$	0.233	0.245	0.253	0.249	0.251
$P_{(20)}$	0.147	0.147	0.155	0.156	0.154
$P_{(30)}$	0.103	0.105	0.111	0.112	0.110

Table 6.8 Results of the effect of weighting on overall similarity scores

Using the same statistical test for comparing the within-subcomponent similarity methods in the last subsection, a two-way ANOVA test is conducted for the between-subcomponent weighting schemes. The results indicate that there are no significant differences of the average precision values among these methods at the 0.05 significance level.

Among the dynamic weight assignment methods, there are two indications from the results that seem to be counter to our expectations. First, the domain-specific information content weighting scheme (Run11) generates an even weaker performance than the domain-independent determination (Run3). Second the equal weighting scheme (Run10) performs surprisingly well compared to the statistically determined weighting schemes. All these seem to indicate that overall similarity determination is not very sensitive to the fluctuation of the inter-part weighting factors. A further observation of the resultant information content values indicates that they do not vary much from each other, as most values tend to fall into the range of 10 to 16, hence the generated normalized weights will tend to be not significantly different from each other.

In general, considered together with the results from the static weight assignments, an inter-part weighting scheme will generate a decent result as long as the headnoun part is not under weighted.

6.6.3 Parsing vs. Non-parsing Model

Our construction of the SIC headings retrieval algorithm is based on the belief that better results would be achieved if the term matching occurs only in the right place—their corresponding functional equivalent parts in a query and a heading. By doing so we try to avoid some noise generated from comparing irrelevant parts. This design approach reflects the principle of alignment-based similarity models whereas heading matching is regarded as an object/concept similarity comparison. Therefore, an appropriate alignment of corresponding parts could ensure a proper matching 'environment'. Thus both query and headings are first parsed to generate a structural and hence semantically indicative representation of the content before a specific matching takes place.

To see how much benefit a parsing model would generate, we tested its counterpart performance with several non-parsing methods. In modeling the non-parsing scheme, it can simply be treated as a special case of the parsing model where the whole query or heading is equivalent to a parsed subcomponent. Therefore, the whole similarity calculation is a many-to-many single terms matching. Similar to our exploration in the subcomponent similarity determination, we can have several methods based on the possible combination of two strategies as 'average of all' and 'best of all' for either query and heading (see table 6.9).

$$Run12\_Sim(t^q_{i-},t^c_{i-}) = \frac{1}{n^q_{i-} * n^c_{i-}} \sum_{k=1}^{n^q_{i-}} \sum_{l=1}^{n^c_{i-}} Sim(t^q_{i\cdot k}, t^c_{i\cdot l})$$

$$Run13\_Sim(t^q_{i-},t^c_{i-}) = \frac{1}{n^q_{i-}} \sum_{k=1}^{n^q_{i-}} \max_{l}[Sim(t^q_{i\cdot k}, t^c_{i\cdot l})]$$

$$Run14\_Sim(t^q_{i-},t^c_{i-}) = \frac{1}{n^c_{i-}} \sum_{l=1}^{n^c_{i-}} \max_{k}[Sim(t^q_{i\cdot k}, t^c_{i\cdot l})]$$

$$Run15\_Sim(t^q_{i-},t^c_{i-}) = Run13\_Sim(t^q_{i-},t^c_{i-}) + Run14\_Sim(t^q_{i-},t^c_{i-})$$

	Run12	Run13	Run14	Run15
P at 0.00	0.572	0.691	0.630	0.762
0.10	0.572	0.691	0.630	0.762
0.20	0.564	0.669	0.610	0.725
0.30	0.547	0.644	0.591	0.707
0.40	0.493	0.624	0.531	0.666
0.50	0.424	0.561	0.452	0.558
0.60	0.408	0.554	0.439	0.545
0.70	0.377	0.508	0.410	0.528
0.80	0.364	0.487	0.398	0.507
0.90	0.361	0.473	0.393	0.491
1.00	0.361	0.473	0.393	0.491
P_{avg}	0.454	0.562	0.494	0.605
$P_{3\text{-point}}$	0.451	0.573	0.487	0.597
R_{avg}	0.347	0.505	0.400	0.442
$P_{(5)}$	0.262	0.345	0.259	0.348
$P_{(10)}$	0.171	0.230	0.175	0.212
$P_{(20)}$	0.105	0.145	0.106	0.131
$P_{(30)}$	0.083	0.110	0.086	0.100

Table 6.9 Results of non-parsing schemes

As expected, the results from non-parsing methods do support the effort made on parsing both query and headings. Each of the combinations in the non-parsing model has a weaker performance than their counterpart in the parsing model (Run12 vs. Run2, Run13 vs. Run3, Run14 vs. Run4, and Run15 vs. Run6). Also, the differences within these non-parsing groups become more apparent than those in the parsing model. This is because the non-parsing comparison is essentially an enlarged subcomponent comparison in the design of the parsing model. Therefore more benefit (from better intra-part determination) as well as more noise would be accumulated in the non-parsing model. One difference between the relative performance of these combinations and those in the parsing model is that we see a fairly significant improvement in the performance of the fusion method (Run15). This would support the similar discovery by Smeaton and Quigly (1996) when they tested the fusion effect on the image caption retrieval. Figure 6.5 depicts the fusion effect in this non-parsing model. From the graph we can observe that this fusion effect becomes more significant in the lower recall points.

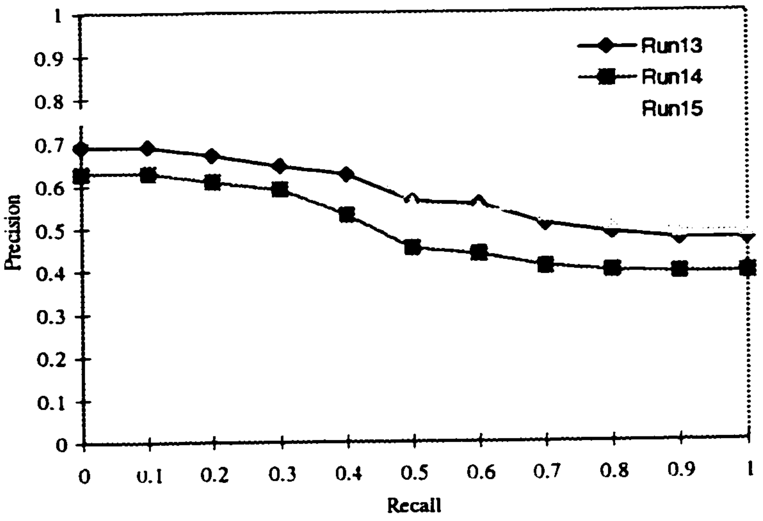


Figure 6.5 The fusion effect in the non-parsing model

6.7 Discussion

In this section we discuss briefly some of the issues that are observed from the design, implementation, and evaluation of this business catalog retrieval subsystem. Though many of the issues are inter-related, we add a short title to each issue to reflect the focus of the discussion.

Advantage of the Similarity-based Retrieval

When performing the same task as SIC heading retrieval, the traditional vector space and semantic similarity approaches come from quite different angles. The former relies exclusively on the surface-level of exact string matching between words, and uses statistical information to weight the various matches to generate a reasonably accurate ranked result. In contrast, the latter goes deeper below the surface, trying to capture the conceptual match between words in a decomposed part, and then arrives at an overall relevance value with the assistance of statistical information.

Trade-off of the Similarity-based Model

While the semantic approach can increase the recall by taking the opportunity of comparing terms that are different in appearance but similar in content, it runs the risks of lowering the precision value. This is reflected by some overall similarity calculations resulting from the addition of individual part similarities. For instance, for the query *Siphon Pump*, example (17) lists some of the top ranking results from the similarity method.

(17) a. 0.692 3312 Rounds, tube--mfg
b. 0.692 3312 Tube rounds--mfg
c. 0.510 3561 Domestic water pumps--mfg
d. 0.510 3561 Pumps, domestic: water or sump--mfg

The results of (17a) and (17b) have much higher scores than (17c) and (17d), which should be ranked higher by human judgment. In WordNet sense definitions, *Siphon* is a kind of *tube*, and the 12th sense of the word *round* has the meaning of a circular rotating mechanism, which is very close to the 1st sense for the word *pump*, which means a mechanical device. Therefore, the added overall similarity value is higher. In contrast, there is a single noun sense for the word *domestic* as home servant, which drags the overall similarity scores lower for (17c) and (17d). Another factor that contributes to this inappropriate overall score is the particular inter-part weight determination. The information content value from SemCor happens to be very high for the word *siphon*. This results in a heavy weight for the modifier part. The major challenge from this example lies in the appropriate way of determining the overall similarity value. Our statistical weighting scheme is of course just one way to approach it.

Need for Fine Alignment

There is more that can be discerned from the above example. The modifier *siphon* in the query indicates the principle or function of the specified subject (*pump*), while the modifier *domestic* (assuming we obtain the correct adjective sense of it) in the heading describes the scope of use for the subject. Therefore, the resulting similarity value would tend to lose its significance if we compare two different *types* of attributes for an object. This again prompts the need for an appropriate feature alignment before comparison. Much more semantic/taxonomic knowledge is required in order to solve the problem. This would be a task that is beyond the consideration of our current algorithm.

Sense Ambiguities

Another more prevailing problem (which can also be revealed from the above example) is the noise from improper senses. In our analysis we do not perform any pre-processes for sense disambiguation. In fact, similarity determination and sense disambiguation seem like circular problems. On the one hand, calculating similarity between word pairs relies on a correctly identified sense for each word. On the other hand, once you find a best match, the sense of a word in the resultant match is typically the correct sense you want to determine.

Other Syntactic Categories

In our decomposed lexical level similarity comparison, the semantic closeness of a pair of words can only be determined if both are nouns. This is due to the restriction in WordNet that the taxonomic hierarchy is only available to nouns. To deal with morphemes other than nouns we either transform them to a related noun by some regular morphological operation or simply perform literal string matching. Therefore, there is a potential loss of information. As for the domain of SIC heading search, this typically reflects the processing of adjectives as they are the second major morpheme in the classification. For instance, the heading in example (18) is ranked very low for the query *spectacles/eyewear frames*. According to our transformation rules, the adjective *ophthalmic* is transformed to noun *ophthalmology,* which means a "branch of medicine concerned with the eye and its diseases". Unfortunately, this transformation does not produce a desired noun for comparison purposes.

(18) 5048 Frames, ophthalmic—wholesale

One relatively simple solution to this is to look into the definition of the morpheme that is not a noun and see if we can find some corresponding noun in it. Many adjective definitions in WordNet list a noun to which a particular sense of the adjective corresponds. For example, the definition of *ophthalmic* in example (18) indicates that it pertains to the first sense of noun *eye*. Although this generation may still not match well with *spectacles* in the query, it is at least getting closer.

Other Semantic Relations

This example brings up another limitation that we have identified in Chapter 4 when we apply this version of WordNet in studying word level semantic similarity. The semantic knowledge we mainly employ for measuring similarity is one (major) type of hierarchical relationship: the hypernym-hyponym (IS-A) relationship. While this type of knowledge has indeed empowered us when tackling major similarity determination problems, there is still the occasion when other types of relations may better address the problem. For example, there is another major type of hierarchical relationship: the meronymy-holonymy (Part-of) relation that WordNet has provided but we have not explored. Also, from our view of object/concept definition in section 2.2, the intent, or attribute of an object would indicate a feature of an object that certain associative relationships can thus be established. For this particular example (18), after we are able to deduce that *ophthalmic* pertains to *eye*, we can discover that an *eye* has *vision* as attribute, and *vision* is associated with *eyeglasses*. Thus a much shorter path linking *ophthalmic* and *spectacles/eyeglasses* is found.

Fair Playing Field

At this stage of prototype implementation, we intend to minimize the use of domain-specific information to enhance the performance. This allows other standard models (i.e. vector space model) to have a fair playing field when compared with our developed similarity model. More importantly, we intended to develop a method that would have a wider applicability. This is why we did not implement steps that might have taken advantage of information from the SIC hierarchical structure. For example, on the one hand we could re-group or re-organize the top rank headings in the resultant set by their 4-digit SIC codes to form a cluster of most relevant headings, and at the same time downgrade some potential 'false drops'. On the other hand, some discarded headings under the current algorithm may be reconsidered for upgrade by their 'siblings' (i.e. having the same SIC codes) that are first entered into the top rankings at the initial retrieval. Many of these will have much to do with the association discovery we discussed in section 6.4.5.

Another potential improvement for the precision criterion is to provide adequate treatment for the word "except" in the headings, as matching parts of the query to words after "except" in the headings will usually result in irrelevant hits. We chose not to do this since we wished to provide an equal playing field for both the proposed similarity approach and the baseline standard vector space model. Complexity, of course, will increase when these elements are considered as negative factors.

Domains of Application

The partial parsing schemes we employ here are weak techniques that require no pre-existing domain-specific knowledge structures. They are simple yet effective in generating the structural regularities that are needed for this task. We see the potential of applying the techniques into other domains. Obviously, the techniques will not be capable of dealing with full-sentence text search and retrieval from larger texts. It is expected to work well when both texts and queries are short when the simple partial parsing scheme is employed. Therefore, potential domains and types of applications are: catalog classification/retrieval, document titles retrieval, image captions retrieval, and the like.

Chapter 7

Conclusion

7.1 Contributions

This thesis targets the problems of language variability from the viewpoint of lexical semantic similarity. We believe that a proper identification of similarity between concepts can contribute significantly in resolving semantic ambiguity in general. We started by defining similarity in a very general sense, and then applied it specifically to various levels of NLP tasks. The complexity of the tasks increases from determining semantic similarity in a single, atomized concept pair, to lexical-semantic disambiguation, to a more complex, multi-layer concept comparison in a real-world IR application.

In Chapter 3, a new formalization of similarity theory is constructed in set-theoretic terms. To lay the foundation for similarity determination, we first presented a general framework in modeling object comparison in set-theoretic notions. Together with a set of certain well-accepted assumptions, this framework has led to a derivation of some general similarity definitions. The derived results can be seen as a generalization of some of the classical definitions of similarity measure. To facilitate a computational effort in determining universal concept similarity, we demonstrated how these established object comparison schemes can be quantified in information-theoretic terms. In general, the new formalization for object comparison is believed to have laid the groundwork for further instantiations using new proposals in modeling object content representation.

The main contribution of Chapter 4 is the introduction of a new method of measuring lexical semantic similarity in a taxonomy. This model enhances the graph distance approach by quantifying the weight of each edge along the shortest path that links two concept nodes in the taxonomy. The word-pair similarity ranking experiment later in the chapter demonstrates that the proposed similarity measure, compared with other related computational models, achieves a result that is closest to humans' performance.

The lexical semantic information derived from taxonomy structure essentially secures a solution to determining the 'commonality' of two objects in information-theoretic fashion, which is crucial to a realization of computational means of resolving universal lexical semantic similarity. This also allows us to develop a unified view of various similarity measures based on taxonomic knowledge, given the background of our general framework about object comparison schemes. Contrary to the common view that 'commonality' dictates similarity, both our theoretic model and empirical verifications have demonstrated that 'difference' is perhaps a better approximation to the similarity measure when object content is measured by its information content.

In Chapter 5, we have demonstrated the utility of applying semantic similarity models in solving some 'intermediate' NLP tasks. A simple word sense disambiguation algorithm is developed which uses the cues from the local contextual information obtained from words surrounding the target ambiguous word. The empirical evidence from the test of tagging all nouns in a running text verifies that the proposed similarity model (i.e. 'difference' model) can generate better performance than other related similarity models in this NLP application.

In Chapter 6, a prototype business catalog retrieval system is designed and implemented. This is a practical application of our proposed semantic similarity method in the area of concept-based information retrieval, a 'final' type of task in the NLP application. In order to provide an appropriate context for a lexical-level similarity comparison, a shallow parsing algorithm is designed to capture both syntactic and semantic information in the catalog headings and queries. The weak technologies employed here require no pre-existing domain specific knowledge structures. Hence the resultant model has appeal to a wider domain of application.

We developed various methods in both decomposing complex linguistic phenomena into single lexicalized items so that previously developed lexical similarity methods can be applied, and aggregating such calculated subcomponent similarities to obtain the overall similarity. For the within-subcomponent parsing and similarity determinations, we designed several algorithms to tackle the syntactic ambiguity problems like compound nouns and complex phrase analysis. For the between-subcomponent similarity determinations, we constructed a framework and proposed two dynamic subcomponent weighting schemes in terms of subcomponent information content value. A prototype system was developed and evaluated against the benchmark of the classical vector space model. The analysis of results indicated that there is a significant improvement over the benchmark (both precision and recall are increased by about 10%).

Finally, there are certain premises that underlie this work. First, the ultimate goal for an IR task is to find documents that are conceptually equivalent to the user's query, therefore an approach that directly applies such a semantic matching scheme would be a significant step towards realizing this goal. Second, an appropriate specification and determination of a similarity/association measurement lies at the heart of this semantic matching scheme. Third, such a measurement should be quantifiable and consistent in treating any object and concept in the universe of linguistics. Fourth, in constructing such a measurement, advantage can be gained by combining both statistical information from unstructured data (e.g. corpora) and linguistic knowledge from a highly structured and organized construct (e.g. MRDs, lexicon, thesaurus).

There has been an unfulfilled expectation that linguistic resources such as lexicons, thesauri or knowledge bases could be exploited effectively in IR tasks (Smeaton 1997). From our implementation of the SIC heading retrieval, though a somewhat special IR task, we have seen such a benefit from the use of a comprehensive lexical taxonomy

(WordNet). This should shed some light on the path of successfully applying various NLP techniques in the IR field.

7.2 Future Research

Several aspects of this research require further work. We briefly discuss each, and, if possible, point toward feasible solutions.

Since humans frequently perceive similarity asymmetrically, it would be interesting to explore an alternative approach to that proposed in Chapter 3, in which 'difference' is defined asymmetrically. Consequently, in Section 4.6, rather than using a distance measure between nodes, we use a difference measure in which the difference between a parent and child in the taxonomy is not necessarily equal to the difference between that child and its parent (for example, a hospital is more similar to a building than a building to a hospitaL)

One issue that has come up repeatedly is the noise that comes from irrelevant senses in the similarity calculation process. As we have discussed sense disambiguation and similarity determination seem to be a circular problem. We need to explore other means to break the 'chain' so that both can benefit.

Due to a constraint of the knowledge source, only the IS-A relation is fully explored in this study. Obviously other types of semantic relations (in particular the associative relations discussed in Section 4.2) also affect the measurement of the semantic relatedness of any two concepts. This can be reflected in assigning a proper link type, $T(c,p)$, factor in formula 4.13 to determine an edge weight in a taxonomy. Experimental investigation of alternative formulas for controlling the influence of node depth and edge density factors might produce interesting insights into how best to estimate edge weights. In aggregating these edge weights, we need to determine the most appropriate link path connecting two end nodes, since different types of links between any two adjacent nodes can then be chosen.

In Chapter 6, when determining the overall similarity score by weighting each subcomponent similarity, we proposed a statistical method based on term frequency information. Although the actual results from various weighting schemes do not seem to vary much, we still believe a better constructed model may represent well the underlying weighting requirement. For example, instead of treating each subcomponent independently for a query, we might see them interact so that each represents a type of feature (attribute or function) of the centerpiece (i.e. the headnoun). In fact, it *is* so from the viewpoint of alignment-based similarity models (section 2.3.3). The problem is how to quantify the importance of each role in contributing to the overall body formed by the entire description.

A similar problem exists in matching product/service descriptions of the EID system: how do we better determine the semantic closeness for each obviously very different dimension in the attribute category space, and conclude with one numerical score after obtaining each dimension's result? It is less difficult if the value of the attribute can be quantified, but more difficult and complicated when only qualitative value data are available, which is often the case in describing product/service characteristics.

The most theoretically challenging work will be either to prove formally our integrated approach (i.e. 'difference' measure) is a manifestation of the transformational model, or to provide an alternative operational vehicle that realizes the transformational model. Achieving this can be seen as a significant breakthrough in similarity research in general.

Appendix A

Psycholinguistic Views of Similarity

As we have identified in Chapters 2 and 3, similarity judgment is a complicated, highly subjective cognitive process. We present here some additional material about similarity and its related concepts, primarily from psycholinguistics and cognitive science.

A.1 Similarity Space

Since the study of similarity is to explore the relationship between objects, which in turn are represented by their underlying component parts as features or attributes, similarity study is then essentially exploring the relationships among those corresponding component parts, or 'respects' so to speak. Therefore in propositional terms, similarity (s) is really a three-place relation $s(a, b, r)$ — a and b are similar in 'respect' r (Medin et al. 1993). There are different types of similarity that can be distinguished depending on the 'respects' in question. One of the most used distinctions is made between *surface* similarity and *deep* similarity. The former is based on the superficial, easily perceived attributes of objects; the latter is based on deep and structural attributes of objects.

This distinction can be better observed when we look at different types of respects in object description and comparison. In one way, respect can be represented as predicate taking argument(s) to express propositions about objects/concepts. There are typically two types of predicates given the characteristics of a predicate's structure: attributes and relations. *Attributes* are predicates taking one argument, and *relations* are predicates taking two or more arguments (Gentner 1983). For instance, *color*(x) is an attribute, while *orbit around*(x, y) is a relation. Attributes are used to state properties of objects; relations express structural correspondences between objects and properties. With this representation, we can more clearly differentiate surface and deep similarity as well as other object comparison related concepts such as analogy, metaphor and simile. Figure A.1, mainly adapted from Gentner and Markman (1997), illustrates the distinction among these concepts.

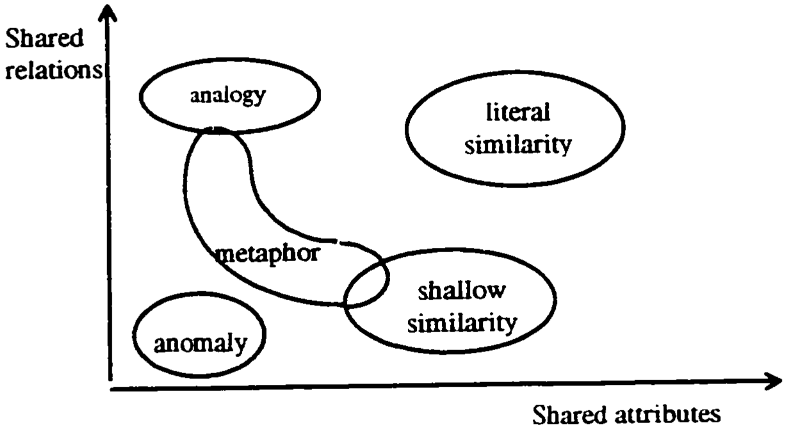


Figure A.1 Similarity space, comparison measures based on share attributes and relations

- Analogy is the process of understanding a novel (target) situation in terms of one (base) that is already familiar. Analogy occurs when comparisons exhibit a high degree of relational similarity with very little attribute similarity. The classic example is the analogy between the atom and the solar system (Gentner 1983). While planets and electrons seldom match in their attributes, they nevertheless have correspondent roles expressed through the relation *revolve*(x, y). Furthermore, deep, cohesive, higher-order relations (predicates taking propositions as arguments) are important determinants of the subjective appeal of an analogy. In this example we can observe a second-order relation: *cause*(*attract*(x, y), *revolve*(y, x)).

- Shallow similarity, or mere appearance match, is in a sense the opposite of analogy. Objects here are compared by the match of certain attributes, not the relations. For example, a round ball and the planet, zebra and jail. Although they can be appealing in a certain sense, their explanatory power is substantially weak, and hence lacks significance. This, however, will lead to certain indications in accessing the development of humans' learning progress.

- Literal similarity requires high degrees of both attributional similarity and relational similarity. An example is a particular star system and our solar system. Typically, the contrast between literal similarity and analogy is a continuum, not a dichotomy (Gentner 1983). This follows from the fact that when two objects overlap in relationships, they tend to be more literally similar to the extent that their attributes also overlap.

- Metaphors are predominantly relational comparisons which are essentially analogies. There are also some with attribute matches. The attributes involved are few but striking. For example, "his lawyers are sharks" (from Glucksberg and Manfredi 1995). Therefore, metaphors can span the regions from high relational comparisons to attribute comparisons.

In general, because there is a direction from the base to the target, analogical and metaphoric comparisons are irreversible (Glucksberg and Manfredi 1995). They will either be anomalous when reversed, or they will entail new grounds for the comparison as in the example "his lawyers are sharks."

A. 2 Similarity and Categorization

When moving from observing and abstracting the relationships from a pair of objects — determining their similarity — to a group of multiple objects, we are actually performing a task called categorization or clustering. The result of categorization is a set of categories where the object similarity within a category is greater than between categories. To a certain extent, the concept of the category is very similar to the concept of the object, if we treat an object's extensional members as another set of objects. Therefore, similar to the classical views about object/concept, traditional models of categories consider that they have rigid boundaries and are defined by necessary and sufficient conditions.

This classical view has been challenged by the so-called prototype model which states that "categories do not have clear-cut boundaries" (Rosch 1978:35). Categories may be defined, "not in terms of what category members all have in common, but in terms of the similarity that category members bear to one another" (Hampton 1993). Central to a category is a goodness-of-example (i.e. prototype) that best exemplifies this category. This expresses the concept of gradedness or typicality for a category. For example, a robin is a very typical bird, whereas penguin is an atypical bird.

Similar to the process of similarity judgment, categorization is also a highly subjective process and involves many interactions with the contextual environment. Lakoff (1987) demonstrated from linguistic and cognitive process points that human categorization is essentially a matter of both human experience and imagination.

A. 3 Beyond the Simple Process

In viewing subjective similarity determination, many approaches regard it as not an isolated construct. Instead it is determined in a given context at a particular stage by a particular person with particular goals. Context, in particular, can bring additional information to the comparison process and then can shift the focus of attention in the process (Tversky 1977, Jurisica 1994).

Appendix B

Designing the Electronic Industrial Directory (EID)

B.1 Introduction

In this Appendix we describe research work that targeted building an Electronic Industrial Directory (EID) — a Yellow Pages-like online business directory for storing and retrieving trade-related business information. One of the three field matching schemes introduced in the EID system — SIC business catalog retrieval — was implemented and evaluated in Chapter 6 as an application of the proposed lexical semantic similarity models.

The organization of this Appendix is as follows. We first introduce the background of the EID project, its motivation, related works, overall design principles, and conceptual modeling. In section B.3, we turn to more detailed descriptions of the directory database design: the database structure, its content representations, and its indexing scheme — the SIC taxonomy. We further discuss the complete retrieval process for the EID system in section B.4. A design consideration about each subsystem is presented. The emphasis is on the concept matching of each field of a directory item with the corresponding part in a query, where various semantic similarity methods can be applied. Finally we summarize the main issues about the EID in section B.5. A prototype implementation of the SIC headings search and various experiments for evaluating its performance is presented in Chapter 6.

B.2 The EID Modeling and Design

B.2.1 Motivation

The ability to carry out business transactions via a computer network is likely to affect greatly the way we conduct business. In this new world of electronic commerce, vendors can set up 'virtual storefronts' on the network to offer direct sales through an electronic channel via an electronic catalog or other more innovative formats. Customers can browse through the catalogs of products/services, identify a desired business counterpart, negotiate via the network, place orders electronically, and obtain the product/service after making the payment electronically.

In the 'virtual storefronts' scenario, a problem that may arise is how to find appropriate 'stores' when there are hundreds of thousands of 'virtual stores' on the network. This creates a demand for a facility that helps the client locate an appropriate store before he/she browses the corresponding product catalog. In a broader sense, the question becomes how to find relevant business information for a participant, considering the immense volume and complexity of the data that exist in the electronic media. As is the case for conventional business, locating an appropriate trade partner is the very first step in pursuing business electronically.

There are various on-line and off-line systems that try to solve this problem by storing business information in databases to help find desired trade partners. However, most of the systems are still in the rudimentary stage. Their emphasis is primarily on the scale and coverage of the data, rather than on their effective retrieval. Many provide retrieval by browsing catalogs. For those who do provide searching functions, the retrieval results are often unsatisfactory due to the use of crude search techniques. In addition, the target users of most of these systems are mostly individual consumers rather than business users.

This raises the question of how to retrieve information effectively, given this application scenario of searching business information from a very large repository. To facilitate research on the subject, we are developing an Electronic Industrial Directory (*EID*) prototype — a database and the means to retrieve information from it (Jiang and Conrath 1996). The fundamental function of the *EID* is to provide intelligent assistance to help business people establish connections. Retrieval effectiveness is emphasized here to guide the design and modeling of this prototype system. This is achieved by employing a concept-based retrieval method to address the semantic matching between the user's query and business item information in the database.

The argument for developing such a system is straightforward. While small-and medium-sized enterprises (SMEs) represent the vast majority of the number of businesses throughout the world and generate most of the new employment opportunities, few have the means or knowledge to get involved in international trade. In particular, they lack the channels to locate potential opportunities. This is a great stumbling block to SMEs' growth, as the world economy has come to the point that participation in it is an important element of a business's economic success.

The other side of the coin is the rapid development and convergence of computer and telecommunication technologies. These are at the stage where electronic commerce is both feasible and desirable. The development of such an *EID* is the first step towards a complete presence of electronic commerce, whereby a business connection is established between buyers and sellers (Conrath 1993, 1994). Furthermore, the electronic brokerage effect would be generated by an *EID* that allows a buyer to compare offerings of many different potential suppliers quickly, conveniently and inexpensively (Malone et al 1987). In a word, the development of an effective *EID* would contribute to the prospect of electronic commerce by providing SMEs with access to vast amounts of company information on line so that they can benefit from seeking and executing trading opportunities.

B.2.2 Other Related Work

Providing users access to business directory type information has been a growing area in commercial database and information retrieval applications. Dun & Bradstreet has perhaps the most comprehensive company database as it contains data on 39 million companies world wide for reference and retrieval. On-line services, such as Compuserv, have Yellow Pages type services for their network community. Many other off-line databases, mostly on CD-ROM, such as Business List-on-Disc by American Business Information Inc., Business Yellow Pages of America by Innotech, MarketPlace Business by Marketplace Information Corp., ProDirect Business by UMI, and CD Business Canada by CD-PowerMedia Productions Inc., provide country-wide business directories (Desmarais 1992, Weide et al 1992). Name search is the major retrieval mechanism for almost all of the operational directory services. Although some offer search by line of business, the user is required to type in an exact industry classification heading in order to find corresponding business information. Furthermore, most of the directories are not particularly relevant for trade, i.e. the content of company profiles contains very little information about an organization's business characteristics, such as product/service descriptions or potential production capabilities.

Several regional Bell operating companies (RBOCs) in the U.S. are co-operating with large computer, cable, or publishing companies to develop large scale on-line Yellow Pages services. Among them, NYNEX has the largest one in terms of volume of business coverage. It is offering a free trial to 7,000 libraries in the U.S. and Canada which would allow them to access over 2.1 million business listings from over 300 Yellow Pages directories from most of its service areas. On-line Yellow Pages services such as this emphasize the scale and coverage of data, and presentation of information (e.g. multimedia business advertisements). RBOCs are taking advantage of their telecommunication infrastructure and already available regional Yellow Pages data.

Electronic commerce has also been introduced to the Internet community, and is experiencing a very rapid growth. This can probably be attributed to the ease of access, relatively low cost, decentralized and free-wheeling nature of the Internet. In particular, the World Wide Web has greatly enhanced the spreading, referencing, and accessing of information, which in turn has expedited the commercial use of the Internet. Electronic Malls (E-malls) or Internet Malls represent well the commercial flavor of net usage. The content of an E-mall usually contains a collection of categorized company listings with the purpose of advertising business products or services. Although efforts have been made to set up sites which contain indexes to other E-malls, most of these are still in the rudimentary stage. Many are implemented as a non-standard hierarchical catalog collection. These perform reasonably well when the volume of data is relatively small. For large-scale systems, some have plugged in a simple full-text search engine designed for the targeted users of the E-mall type electronic directories—individual end consumers. At the time of this writing, there have been no fully functional ubiquitous directory services that customers can rely on to search by line of business, or by product/service.

In the attempt to address more directly the fundamental issue of information retrieval (IR), concept-based models have been introduced to overcome the inherited problem of traditional word-based approaches which manipulate textual data at the lexical level (Chen et al 1993, Ginsberg 1993, Ledwith 1988, Lee et al 1993, Rada et al 1989, Voorhees 1994). Various techniques from artificial intelligence and natural language processing (NLP) have been employed to enhance system capability. The result is a system that behaves more 'intelligently' by means of greater 'understanding' of a user's query and providing more relevant and complete results. This is mainly achieved by the use of extra space for storing the knowledge base, preprocessing the constructions and taking sufficient time to navigate the search space. Commercial tools, such as Topics, PLS, ConSerach, Thunderstone, are typical of this trend, although none have been directly applied to business directory retrieval.

The proposed EID is a semantic/concept-based approach that targets a niche in this opportunity space. The research faces similar problems of managing and retrieving information from a very large and complex text repository. Our emphasis is on the importance of effective hits in database search, rather than simply plugging a full-text search engine. This requires a good understanding of the characteristics of business information, both structure and content, as well as designing and building an intelligent front end to facilitate search and retrieval.

B.2.3 The EID Design Principles

Before we present the detailed modeling and design of the proposed EID system, we outline some design principles that direct our research methodology. We believe these concepts and principles are also applicable to other similar types of intelligent retrieval applications.

First, the database should be ordered according to some conceptual classification scheme and presented in a systematic structure. A short semantic notation is needed to label the content of each item stored in the database. This is analogous to the classification number assigned to each item in a typical library catalog system. The adding of a taxonomy would act as an augmented semantic network that will facilitate the conceptual searching scheme.

The proposed business listing structure should be as open and as extensive as possible. That way new company listings with various structures can be easily converted and categorized into the existing structure. Selected structures should be able to cover the domain of typical business trading information.

We emphasize the ease of use of the system, as most users would have limited experience with sophisticated computer systems. This requires that the system behave more 'intelligently' to accomplish tasks under various scenarios. The designs of query/profile input and query reformulation need to provide the user with a simple and straightforward interface. Appropriate intelligent assistance and feedback to the user are expected. It is desirable to let the system do as much work as it can while leaving the user a certain degree of flexibility in controlling the interactive process.

In query processing, rather than simply utilizing a full-text retrieval engine, a concept-based retrieval approach is pursued. Going to the fundamentals of information retrieval, what we are seeking is a semantic equivalence between a user's query and the retrieved item. In order to achieve this, consideration is given to the design of a data structure that reflects the conceptual relationships among the stored items. A pre-coordinate model is sought to facilitate this organization. The retrieval process is to narrow the target area by searching the concept search space. Hence emphasis is on the importance of obtaining a few effective hits in a database search, i.e. a well-balanced precision and recall performance.

In the line of semantic matching, a method of measuring conceptual similarity is required. This will then allow the system to generate ranked results. It can be implemented by determining the conceptual distance between items in the search space. Furthermore, the system should be able to keep track of associative information (i.e. the potential association of two items given a certain aspect of relatedness among items' subcomponents information) when each new item is entered.

A browsing function is expected to complement the searching method for such a large and complex database. When necessary, the system should enable a user to navigate the semantic search space to clarify the searching position and find associative information as well.

B.2.4 The EID Conceptual Modeling

The two major functions of the EID are the storage and retrieval of business related information. A schematic representation of the architecture is shown in Figure B.1. The left side of the diagram corresponds to the information gathering procedure, and the right to the subsequent retrieval. Starting with input, there are three ways one might expect to obtain data about an organization. One would be by means of direct input, perhaps using an interactive on-line system, either by the company itself or via a designated third party. Another possibility would be batch input of company profiles from other media or databases. While the data may not provide as complete a description as desired, this would economize both time and effort. Third, the system might automatically search the network (e.g. Internet) to fetch a company profile from a specific company server (e.g. Web pages). The last two methods would need a conversion tool to translate the format of the original data to that required by the EID.

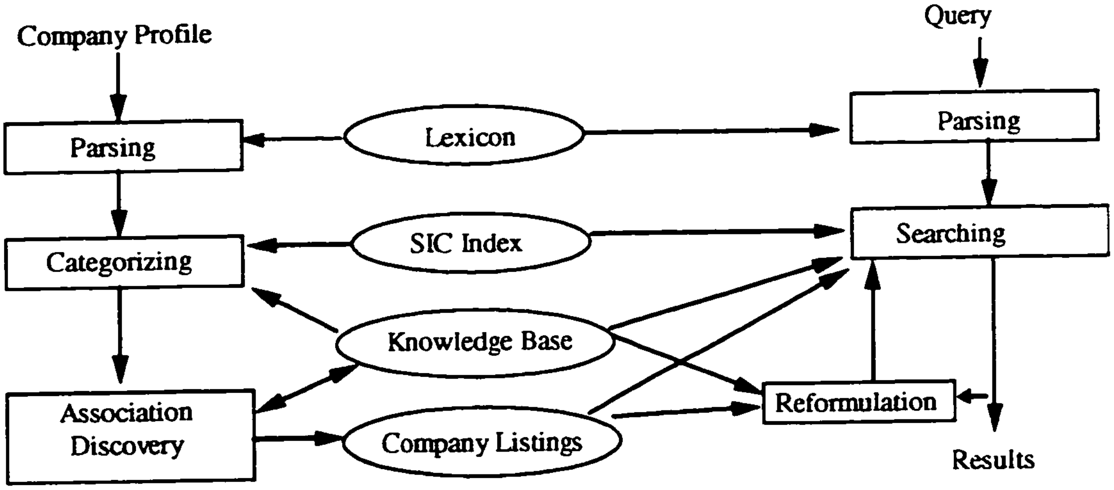


Figure B.1 Architecture of EID

Company data gathering and organization. A company's profile includes basic information such as name, address, and telephone number, as well as the nature of its business activities and product/service descriptions. After the information is parsed for grammatical processing, the profile is categorized based on the Standard Industrial Classification (SIC) scheme and stored under the corresponding business category. During the categorizing process, the system would recognize the characteristics of this specific business and establish association link information which is stored in the knowledge base to aid later retrieval.

Company data retrieval. The retrieval process begins when the user presents his/her product/service needs in a format which provides a straightforward query generation. Similar to company profile parsing, the query is then parsed for grammatical processing. Searching, which is the key to the retrieval process, follows. It involves applying a concept-based matching procedure (as yet unspecified) that would filter the data contained in the directory in response to the features specified in the query. In many cases the initial results may not be satisfactory. Thus query reformulation would be undertaken to adjust the constraints in the query until the outcome is perceived as satisfactory.

B.3 The EID Database Design

B.3.1 Directory Content Representation

The main part of the Directory will be a text database that contains pertinent data about businesses. For each company, for which there is a record, a list of fields is presented. Each field represents an attribute of a company. The overall structure is similar to that of a typical relational database, only the data types for most of the fields are in a textual format.

Rather than specifying exhaustively all the possible attributes in the database structure, we define only a few important fields that collectively provide a reasonable coverage of information for a company engaging in trade, while individually representing a significant characteristic. This kind of balance conforms to the openness and extensiveness principles of the database design.

- Company name.
- Company address.
- Electronic contact.
- Standard Industrial Classification (SIC) headings with code.
- Product/Service names.
- Standard Classification of Goods (SCG) headings with code.
- Product/Service descriptions.
- Business advertisement.
- Demographic factors.

The Product/Service descriptions field is expected to be a relatively long textual field containing detailed descriptive information such as product/service specifications, price, optional features, etc.

Demographic factors include data such as size of business, annual sales, business history (typically years of business in current category), assets, etc.

Using the conventions of the vector space model, company profiles can be represented as vectors in hyper-space (Salton and McGill 1983). A distance metric which measures the conceptual proximity of vectors to each other is defined over that space. This is very useful in providing ranked results to the user.

Each field consists of one or more subcomponents or parts that further identify the whole. The subcomponent part is a smaller syntactic unit as opposed to a complete sentence or phrase of the encompassing field. Depending on the grammatical processing, such decomposed parts can contain a single word or several words that represent a relatively independent concept. We will detail this in section B.4.4. Formally, a field can be represented as a subcomponent vector:

$$F_i^c = \left\langle t_{ij}^c \right\rangle,$$

where t_{ij}^c is the jth part in the field F_i^c . The superscript c indicates that F_i^c is a company field, as opposed to a query field described later. Thus a company profile can be represented as a set of fields:

$$C = \{F_i^c\}.$$

B.3.2 Indexing Scheme – the Standard Industrial Classification (SIC)

As mentioned in the design principles, we need a semantic notion to label each item (company listing) in the database, so that the whole database can be structured in a meaningful manner. From the pre-coordinate system point of view, this is actually the preprocess of finding a proper indexing or classification scheme to categorize the source data. After an extensive investigation, the Standard Industrial Classification (SIC) appears to be the best alternative for such a business taxonomy. In fact, the SIC is also widely used in many other Yellow Pages-like systems. Thus we have adopted it to form the framework (index) of the main directory database, especially since the SIC scheme also provides a hierarchy which is useful in supporting database navigation and browsing.

There are several official SIC standards, such as those developed by the United Nations (United Nations 1990), the U.S. government (U.S. 1987), and Canada (Statistics Canada 1980). All of these standards use a four-digit coding scheme to form a hierarchical classification structure. Among them the U.S. government standard has the most comprehensive coverage in term of the total number of headings. From now on, if not specified elsewhere, the discussion will be focused on the U.S. version only.

B.3.2.1 Purposes and Principles of the SIC

The SIC was developed for use in the classification of establishments by type of activity in which they are engaged; for purposes of facilitating the collection, tabulation, presentation, and analysis of data relating to establishments; and for promoting uniformity and comparability in the presentation of statistical data describing the economy. The *establishment* here refers to an economic unit that produces goods or services (US 1987, p.699). Each establishment is classified according to its primary activity. Primary activity is determined by identifying the predominant product(s) produced or service rendered. The establishment differs from a classification for enterprises (companies) or products. An enterprise consists of all establishments having more than 50 percent common direct or indirect ownership. Other classifications have been developed for use in the classification of commodities or products (e.g. SCG) and also for occupations.

B.3.2.2 Structure of the SIC

The structure of the classification makes it possible to classify and analyze establishment data on a division, a two-digit major group, a three-digit industry group, or a four-digit industry code basis, according to the level of industrial detail considered most appropriate. Additional subdivisions within specific four-digit industries may be allowed to address further detailed specific needs for an agency.

Hence a hierarchy of the SIC code (with four digits) looks like the following:

- Division (A--K)
    - Major Group (01-99)
        - Industrial Group (XX1-XX9)
            - Industrial Class (XXX1-XXX9)

Under each division, there are number of major groups varying from 3 (Division C Construction) to 20 (Division D Manufacturing). The three-digit industrial group is an expansion of each major group. Similarly, the four-digit industrial class is an expansion of its upper class industrial group. It should be noted that the digit '9' appearing in the third- or fourth-digit position of the classification code usually designates miscellaneous three-digit industry groups or four-digit industries covering establishments not elsewhere classified. These residual classifications do not usually constitute homogeneous primary activity groups; for purposes of classification they are grouped together and treated as a separate industry to retain the homogeneity of the other industries in the group.

A snapshot of the classification hierarchy is given in Figure B.2, where headings are prefixed with their corresponding codes.

```
...
- Division D  Manufacturing
        - 20 Food and Kindred Products
                - 201 Meat Products
                        - 2011 Meat Packing Plants
                        - 2013 Sausages and other Prepared Meat Products
                        - 1015 Poultry Slaughtering and Processing
                - 202 Dairy Products
                        - 2021 Creamery Butter
                        - 2022 Natural, Processed, and Imitation Cheese
                        - 2023 Dry, Condensed, and Evaporated Dairy Products
                        - 2024 Ice Cream and Frozen Desserts
                        - 2026 Fluid Milk
                ...
        - 21 Tobacco Products
                - 211 Cigarettes
                ...
        ...
        - 39 Miscellaneous Manufacturing Industries
                - 391 Jewelry, Silverware, and Plated Ware
                ...
                - 399 Miscellaneous Manufacturing Industries
                        - 3991Brooms and Brushes
                        ...
                        - 3999 Manufacturing Industries, not Elsewhere Classified
```

Figure B.2 A snapshot of the SIC system (U.S. 1987)

Under the bottom (i.e. 4th) level of an industrial class, there is usually a further listing of specified subheadings that belong to it. They represent a decomposition of that specific industrial class. For example, there are 26 subdivisions under the industrial class *2011 Meat Packing Plants*, which are listed in Figure B.3. We will return to the use of such subheadings shortly.

Bacon, slab and sliced
Beef
Blood meal
Boxed beef
Canned meats, except baby foods and animal feeds
Cattle slaughtering plants
Corned beef
Cured meats
Dried meats
Frankfurters, except poultry
Hams, except poultry
Hides and skins, cured or uncured
Hog slaughtering plants
Horsemeat for human consumption
Lard
Luncheon meat, except poultry
Meat extracts
Meat packing plants
Meat
Mutton
Pork
Sausages
Sheep slaughtering plants
Slaughtering plants: except animals not for human consumption
Variety meats edible organs
Veal

Figure B.3 A list of subdivisions for industrial class "2011 Meat Packing Plants"

Unfortunately, the official SIC codes are not detailed enough for many users who might have specific products or services in mind. For example, the current U.S. SIC codes only have roughly a total of 1,000 industry entries (up to the fourth digit code). An expansion of the SIC code was undertaken by Dun & Bradstreet using the U.S. government SIC as the basis (Kresge 1990). Their expanded version (called SIC 2+2) uses up to 8 digits (two more hierarchical levels). This increased the actual number of industry classes to about 15,000. We had planned to adopt this expanded version to classify the SIC field for our directory database. However, due to the lack of availability, at present and throughout the discussions in this thesis, the SIC headings studied refer to the U.S. government version (U.S. 1987) only. Bearing in mind of embedding a classification that is as comprehensive and specific as possible, we take full advantage of the available official SIC classification, i.e. we have included all the subheadings under each four-digit industrial class. This expansion increased the total number of fourth level business headings to over 17,800 which is even larger than the Dun & Bradstreet's expanded version SIC 2+2. In fact, their two more levels expansion was undertaken on the consideration of those subheadings that already exist under each fourth level heading in the U.S. government version. Since we are primarily interested in the coverage of a classification at this point, this expansion is sufficient for our prototype work and the performance evaluation discussed in Chapter 6.

B.3.2.3 The future of the SIC -- the NAICS

The SIC system was developed in the 1930s as a means of collecting and summarizing statistical data on industries in the United States. During this time, the U.S. economy was primarily manufacturing-oriented. It has been revised about every ten years to reflect the changes of the industries and the economy. The last major revision to the SIC occurred in 1987. With the growth of the global economy, emerging technologies, and industries such as the service sector in particular, in 1992, the US Office of Management and Budget (OMB) established the Economic Classification Policy Committee (ECPC) to study ways to improve the SIC system. In 1994, the ECPC launched a campaign to develop a new classification system. The result is the NAICS--The North American Industry Classification System, a classification system which creates one uniform system for classifying industries across the continent. As a joint effort of the ECPC, Statistics Canada and Mexico's INEGI, NAICS was implemented in 1997 and will be updated every five years.

Based on a production-oriented conceptual framework, NAICS reflects the increasingly service-oriented nature of the economy and recognizes new and emerging technologies. Like the SIC, NAICS has a hierarchical classification structure but with up to 6 digit codes:

XX	Industry Sector
XXX	Industry Sub-sector
XXXX	Industry Group
XXXXX	Industry
XXXXXX	National Industry (not always used)

As we do not have access to this newly developed NACIS codes, we will keep the SIC codes in our prototype development. Since there is no major structural shift caused by the development of NACIS, it is expected that there will not be any major affect when upgrading to the NACIS codes in the future.

B.3.3 Overall Directory Structure

Figure B.4 depicts a schematic relationship between the SIC hierarchy and the main company listings in our database design. Notice that companies are usually linked to the lowest level of an industry class in the SIC hierarchy, since the scope of an enterprise is usually very focused and limited. A company may also be linked to several industry classes as it may have more than one establishment.

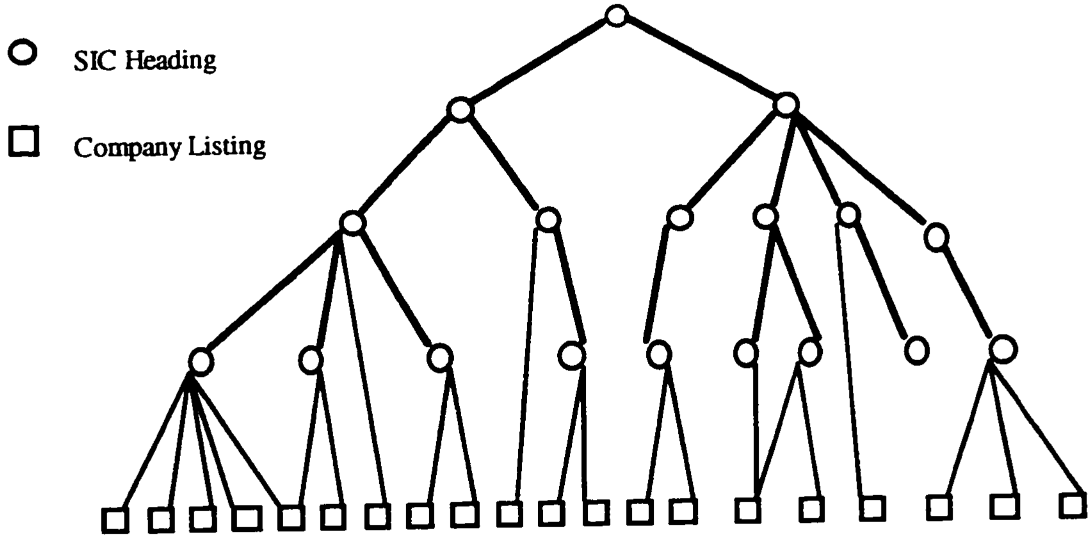


Figure B.4 The relationship between SIC headings and company listings

B.4 The EID Retrieval Modeling

To develop appropriate retrieval processes we need to understand the needs of people who seek suppliers and/or purchasers. Most likely the seeker has a reasonably well defined idea (rough specifications) of the product or service sought. If the person knows the names of the potential suppliers (or purchasers) the problem is trivial. If one does not have specific companies in mind, the next step would be to match what one seeks with a category in one of the indexes, such as the SIC or SCG codes (Statistics Canada 1993), analogous to the way one uses subject headings in the Yellow Pages. Again, if one knows the appropriate code(s) or the correct corresponding heading terms, the most likely problem is to ensure that the number of alternatives is not overwhelming. This might best be done by adding additional constraints such as geographic location, size, number of years in business and the like.

The major problem arises when the user is not sure what code to use or when there is a need or desire to scan the product/service description part of the database. For this, one needs help. We will describe our approach by discussing the complete searching procedure: query interface design, parsing, processing and reformulation.

B.4.1 Query Interface

Query By Example (QBE) or form-based query is a commonly used query format (e.g. in library systems or bibliography retrieval). This is particularly useful for a structured (e.g. fielded) textual databases. We implement our query format in a manner similar to the QBE method. Users key in the desired information needs in the corresponding field. We intend to start by setting up just a few fields, rather than a complete set defined in the directory database. This serves two purposes: first it makes it easier for a novice user to express his information needs without much understanding of a specific field definition in the database. Second, the system has more flexibility in grouping (sometimes cross-referencing) similar database fields to match a query field to assure better retrieval effectiveness. Initially the following three query fields are presented to the user:

- Industry/business category (corresponding to the SIC headings)
- Product/service requirement
- Geographical location preference

As the specified terms in each field represent a desired property/restriction on the overall requirement, the corresponding field(s) in the retrieved company listings may not necessarily contain the exact text terms, but rather a content equivalence or satisfaction. For instance, a specification of *New England* in the geographic preference field is allowed although in the actual profile data no such term would appear in the address field of company listings.

Formally, similar to the company profile representation, each field in a query will be expressed as a vector:

$$F_i^q = \langle t_{ij}^q \rangle ,$$

where subscript i might be an industry/business category, product/service requirement, or geographical preference. A query is then a set of its fields:

$$Q = \{F_i^q\}.$$

B.4.2 Query Preprocessing

When a user submits a query it is first parsed by some simple morphological and lexical processing. Since we allow natural language-like query input, which involves no Boolean operators or other function specific symbols, this step basically involves morphological analysis such as stemming, and some lexical processing such as the expansion of abbreviations via a dictionary look-up, and stop-word removal. There may be some semantic issues involved that would make use of a general thesaurus. This would result in several possible semantically equivalent queries.

B.4.3 Query Processing

A simple algorithm is presented for the initial query processing. The idea is to filter the database with the field specifications in a specific order. Since company listings are categorized to the corresponding SIC headings, the SIC field should be filtered first. We add other constraints to the initial resulting set to increase retrieval precision. Product/Service Name and Business Description fields are filtered next. Last, the

Company Address field is filtered. We assume the initial complete set of company listings is C_0, and the complete set of SIC headings is H_0.

Step 1. Filter set H_0 to generate set H_1, where H_1 matches field F_i^q (i=industry/business category). Generate C_1 from C_0 where the field F_i^c in C_1 equals H_1. If either $H_1 = \phi$ or $C_1 = \phi$, then $C_1 = C_0$, go to step 3.

Step 2. If the number of companies in set C_1 is lower than a threshold, then $C_2 = C_1$, go to step 4.

Step 3. Filter set C_1 on F_i^c (i = product/service name, product/service descriptions, SCG headings) fields, to produce set C_2, where F_i^c matches $F_{i'}^q$ (i' = product/service requirement). Rank the resulting set based on similarity measures.

Step 4. Filter set C_2 on field F_i^c (i = geographical location), where F_i^c matches F_i^q. This results in set C_3.

Step 5. Present set C_3 of the business listings to the user in ranked order, with the associated score (which led to the ranking) attached to each item on the list.

The effectiveness of retrieval is emphasized. In step 1, when no industrial class is found (i.e. $H_1 = \phi$, which means an inappropriate specified industry category in the query), or no company listing is found under the matched SIC heading (i.e. $H_1 \neq \phi$ and $C_1 = \phi$), the searcher has another chance to obtain related information as the required information may be found from searching other business function related fields. In step 3 we match the product/service attributes field in the query with three other related fields in the company profile data. This provides a certain degree of flexibility in the filtering process. The matching and ranking operations among these steps are based on the conceptual relevance calculation described in the following subsection.

B.4.4 Methods for Conceptual Mapping

For each of the above filtering steps there is a searching/matching involved in a specific field. These comparisons are not literal string matching, but rather a content/semantics matching. We will employ our semantic similarity model developed in the previous chapters, as well as others techniques from IR and NLP to develop proper matching schemes to address these issues. The essence of these is to try to analyze and 'understand' both the user's query and the company profile before a comparison is pursued. There will be different matching processes (hence different similarity metrics) for each field filtering. The following subsections present a brief discussion of each field matching scheme. More detailed implementation for the business/industry category field matching will be presented in Chapter 6 as an example of a specific implementation.

B.4.4.1 Industry/Business Identification

The matching operation for the Industry/Business category involves identifying some industry heading(s) in the SIC hierarchy that is/are closely related to the user's input description. This step is a relatively easy one among the three proposed field matching requirements, as the representations of both user's input of an intended business category and SIC heading are in a somewhat simple natural language format. Often they are concise and described in a few words in a noun phrase-like format. This makes it easier in the parsing step in order to derive a set of relatively independent subcomponents. The hard part is to correctly interpret the user's terms and expressions as well as the SIC headings. A design for this step includes the following:

1. Parse the business category description to generate its concept structure,
2. Locate the corresponding SIC heading(s) in the hierarchy,
3. Calculate the relevance of the selected heading(s),
4. Expand the concepts represented in the query if results are not satisfying, repeat from step 2.

The above procedure describes a conceptual approach to matching an industry class with the help of both lexical and semantic knowledge. It first uses grammatical rules to analyze the query term and break it down to the minimal semantic units, then performs conceptual matching for each semantic unit against the counterpart from a potential relevant heading. The logical structure of a taxonomy is used for exploring the conceptual connections. Through tracing the relationships among concepts/classes this algorithm is able to draw out a user's implicit yet conceptually relevant query term(s) based on the explicit term concept it contains.

It should be pointed out that this is a rather coarse description of the solution. The specific implementation is discussed in Chapter 6.

Calculate the relevance of the selected heading(s). We present a formula to calculate the conceptual relevance between the user entered business category and the selected SIC headings. A similarity function is used to represent this semantic relevance:

$$Sim(F_i^c, F_i^q) = \sum w_{ij} Sim(t_{ij}^c, t_{ij}^q),$$

where subscript i denotes a business category matching, subscript j indicates the decomposed part of an input query, w_{ij} is the weight factor for each decomposed part j and is to be normalized. Weighting can be fixed for each subcomponent part for all queries, or can be dynamically determined based on the relationship of the subcomponents for a specific query. Again, this, together with the subcomponents similarity determination, will be explored in our later experiments.

B.4.4.2 Product/Service Comparison

As we allow users to express their product/service needs in a rather free-form manner, the function of matching product/service descriptions would require more syntactic and semantic processing than that required by the SIC field. This is also because the content of such descriptions usually contains textual terms with embedded numerical values representing more complicated, detailed and specific requirements. Several techniques are required to fulfill the task of matching a company's products/services with a desired information need.

Similar to the previous parsing methods used in solving industry headings matching, we can break down the product/service description field into smaller and pertinent semantic units. Each semantic unit represents one dimension (attribute) in describing the characteristics of a specific product/service. Comparison is then performed on the values of each matched attribute.

Since a semantic unit may not necessarily correspond to a syntactic unit (e.g. a sentence), there is a problem of how to properly identify these attribute descriptions that are embedded in a natural language statement. One simple approach is to identify the term (called *attribute identifier*) that characterizes the attribute first, and then pick up a window region of text within which the terms representing the value of the attribute identifier would reside. Table B.1 lists in alphabetical order of some common attribute identifiers used in product/service descriptions for trade purpose.

Application (Use for)	Preference
Color	Price (Target price)
Commodity	Purity
Composition	Quality
Delivery	Quantity
Destination	Shipping cost
Moisture	Shipping terms
Origin of goods	Size
Packing	Validity
Payment	Weight

Table B.1 Common Categories of Product/Service Attributes

Given an identified window region of an attribute category, the following issues need to be considered when determination of its value is performed:

- The value of an attribute that specifies a characteristic of a company product/service is typically textual, often embedded with numeric quantities, ranges, or special symbols to represent product specifications or configuration. Examples are dates, part numbers, nature of currency and chemical formulae.

- Units of measure may also be different between directory company data and the query terms. This is especially true for attributes containing range values (e.g. temperature, weight, volume, etc.)
- A variety of abbreviations of terminology and measuring units is often encountered in both query and directory data for product/service specifications.

Specific functions are to be built to implement the recognition of numeric information in the text of company data in order to compare it with a user's inputs, the latter of which is also more or less represented in a natural language fashion. A planned rule-based expansion of tne equivalency relationship to link natural language terms to numerical values will alleviate the problems arising from various data types.

B.4.4.3 Geographical Preference Matching

Satisfying a geographical preference requires that the system's database contains a fair amount of geographic location information, which then forms a hierarchical network similar to the SIC headings structure. As a case in point, if a user specifies a preference in terms of a region (e.g. *North America*) rather than a city, province/state, or country, the system should be able to recognize this and convert it to some equivalent geographic area which contains certain disjunctive classified areas (e.g. cities, provinces/states, or countries) that are the next lower level to the entered region in the geographical hierarchy.

B.4.4.4 Other Constraint Requirements

Some users may be interested in identifying other fields as part of the query constraints. For example, the anticipated number of resulting companies that satisfy the restrictions set for the three proposed fields may be large, or the user may have special requirements that pertain to other non-business function related attributes (e.g. years in business, size, related expertise.). These operations can be implemented in a fashion similar to the above semantic filtering techniques.

B.4.5 Query Reformulation

The overall retrieval process works in two steps, using a combination of searching and browsing through SIC headings and business listings. The system first locates potential industry classes (and the subordinate company listings) by searching against the initial query. Then a fine-tuning procedure is pursued via query reformulation by allowing users to browse through a portion of the SIC hierarchy that contains the located SIC class node(s), and to further refine the query.

Users can refine the query given the retrieval results and the current location of the industry node(s) in the SIC hierarchy (see Figure B.5). A graphical form will be presented to the user to display the relevant portion of the hierarchy. Treating the SIC hierarchy as a special case of a thesaurus network, several possible directions can be pursued for query refinement (see Chen and Dhar 1990, Chen et al 1993, Kimoto and Iwadera 1990).

BT The user can go up one level in the hierarchy to obtain a *broader term* (BT) view, where company listings under the parent node are chosen. This would be the case when the products/services provided by the retrieved company listings are too specific and the user wants to find some broader capacity.

NT The user can go down one level for a *narrower term* (NT) if a more specific industry class is preferred. In this case, the user can specify which child node to pursue given the alternative subclasses.

RT Often, finding *related term* (RT) information is the desired alternative. One way to find an RT is to ask the user to select a sibling node at the same level as the current node in the SIC hierarchy. In addition, a class node or a company may have a related party in a distant branch with which it was linked in the profile categorization process through association discovery (see Figure B.5 and the next section). This linking may also be treated as an RT relation. In general, the resulting set of this sort of semi-automatic query reformulation can be controlled by the user as he/she navigates through the classification network.

The broadening/narrowing procedure may not produce the desired result due to the local density of the SIC hierarchy (i.e. the number of branches spanning out from a specific node) and the local density of actual data (i.e. the number of companies belonging to that class node). The control of navigating through the hierarchy can be given directly to the user, or it can be undertaken by means of automatic density determination (e.g. an indication of the number of companies on a specific node and its neighbors) where the system determines the direction and path length of the navigation.

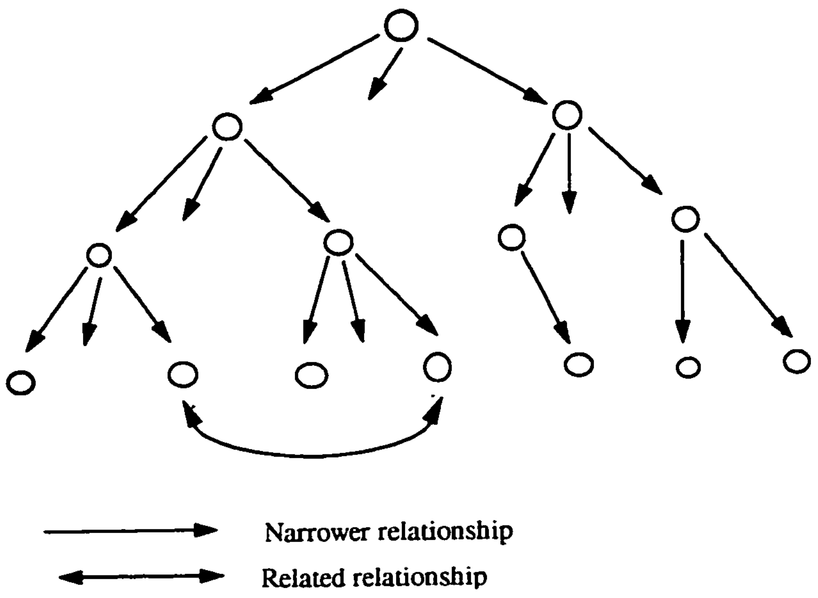


Figure B.5 The structure of the SIC hierarchy

B.4.5 Association Discovery

Another dimension of intelligent retrieval assistance is through discovering special associations among industry classes or among company profile listings. This is very useful because there are industries which are placed into different classes in the SIC hierarchy, but which are actually closely related from a practical business point of view, for example producers and their suppliers. The 'industry type' in SIC headings can also indicate associations among industries, for instance two industries where one is producing a tool and the other repairing it. Moreover, two companies in different categories may have considerable overlap in their product/service characteristics. Identifying these industries or companies as associated ones (e.g., an RT relation in the SIC hierarchy) would help users in further query reformulation. We intend to have both static and dynamic association discovery performed by the system.

Static association discovery can be implemented similarly to the conceptual matching schemes that are executed in the query processing step. By applying this to our SIC hierarchy we can determine the distance between any two nodes in the SIC hierarchy for both industry classes and the subordinated company listings. Once the calculated association is over some predefined threshold, an RT link can be set up between the two nodes. The comparison of two company profiles will cover all fields, but higher weights will be given to the SIC heading and product/service description fields.

Dynamic association discovery is a learning process whereby the system itself is able to accumulate knowledge through the process of running the directory retrieval. Some of this knowledge can be obtained by keeping track of the users' behavior and feedback to the searching process and the resultant set. For example, if two industries or companies are often selected by users at the same time, that suggests they are related. An association link (RT) can thus be established. Other learning rules may be generated as we test the prototype when a wide spectrum of users are selected with the belief that the system can become more 'intelligent' through this learning process.

B.5 Summary

In this Appendix we described a practical application of the proposed semantic similarity model in the area of business directory retrieval. We presented a conceptual model of the *EID* and a design approach for each of its subsystems. The objective of the system is to provide easy access and effective retrieval of company listing information for an unsophisticated user seeking a potential supplier/purchaser. The major retrieval service offered is a search by lines of business reflecting product/service requirements, enhanced by constraints indicating preferences as to location, size, experience, and the like. The system provides maximal ease of use as no special knowledge, such as industrial classification, is required by the users. Visual navigation of the search space is supported via an interactive searching process. Among the areas studied, intelligent retrieval based on semantic similarity matching is emphasized to differentiate our work from existing commercial products/services which stress the scale of the database rather than effective retrieval from it. Furthermore, the net result is likely to involve a database far richer than what one currently finds in the Yellow Pagcs, which in turn suggests that more research will need to be done in both coding an organization's characteristics and the structuring of descriptive information.

Our approach to concept-based retrieval came with this rationale: we believe a pre-coordinate system (e.g. using the SIC hierarchy as an indexing framework) will provide a basic 'understanding' of the directory's storage content. This will greatly benefit the retrieval process in terms of semantic searching. When dealing with this semantic matching, rather than exclusively relying on a universally comprehensive knowledge base (i.e. WordNet), we resort to corpus statistics to assist in calculating the semantic similarity between concepts and classes. This combination of knowledge base and corpus statistics methods gives us the advantage of obtaining effective hits from large business data retrievals.

Appendix C

Test Set for SIC Headings Retrieval

The following is a complete list of all the 95 valid entries that form the test set (queries) for the experiment of SIC headings retrieval conducted in Chapter 6. For each entry, all the relevant headings (answer set) are also provided. All the entries are sampled from a daily international trade mailing list called Trade-L. The original format (e.g. cases) is mostly preserved here unless some necessary modifications (e.g. spelling errors) are required.

1 Aluminum Containers
3497 Containers, foil: for bakery goods and frozen goods--mfg
3497 Foil containers for bakery goods and frozen foods, except bags and liners-- mfg
3411 Cans, aluminum--mfg
3411 Cans, metal--mfg
3411 Containers, metal: food, milk, oil, beer, general line--mfg
3411 Food containers, metal--mfg
3365 Utensils, cast aluminum--mfg

2 spectacles/eyewear frames
3851 Frames and parts, eyeglass and spectacle--mfg
3851 Eyeglasses, lenses, and frames--mfg
5048 Frames, ophthalmic--wholesale
3851 Mountings, eyeglass and spectacle--mfg

3 AIR FILTRATION SYSTEMS
3564 Air cleaning systems--mfg
3564 Air purification and dust collection equipment--mfg
3564 Filters, air: for furnaces and air-conditioning equipment--mfg

4 Athletic Shoes
3149 Athletic shoes, except rubber--mfg
5139 Athletic footwear--wholesale

5 Food Additive
5169 Food additives, chemical--wholesale

6 fabric conditioner
2842 Fabric softeners--mfg
2843 Softeners (textile assistants)--mfg

7 PETROLEUM RESIN
2821 Petroleum polymer resins--mfg
2821 Resins, synthetic--mfg
5162 Resins, synthetic: except rubber--wholesale

8 canned chickens
2015 Chickens, processed: fresh, frozen, canned, or cooked--mfg
2015 Poultry, processed: fresh, frozen, canned, or cooked--mfg
2032 Chicken broth and soup, canned--mfg
2011 Canned meats, except baby foods and animal feeds--mitse--mfg
2013 Canned meats, except baby foods and animal feeds--mfpm--mfg
2015 Turkeys, processed: fresh, frozen, canned, or cooked--mfg
2015 Ducks, processed: fresh, frozen, canned, or cooked--mfg
2015 Geese, processed: fresh, frozen, canned, or cooked--mfg

9 Electric Heaters/Pumps
3585 Heat pumps, electric--mfg
3569 Heaters, swimming pool: electric--mfg
3634 Electric space heaters--mfg
3634 Heaters, immersion: household--electric--mfg
3634 Heaters, space: electric--mfg
3634 Immersion heaters, household: electric--mfg
3634 Room heaters, space: electric--mfg

10 Fresh and Frozen Seafood
2092 Seafoods, fresh and frozen--mfg
2092 Fish: fresh and frozen, prepared--mfg
2092 Frozen fish, packaged--mfg
2092 Frozen prepared fish--mfg
2092 Shellfish, fresh and frozen--mfg
2092 Shellfish, fresh: shucked, picked, or packed--mfg
5142 Seafoods, frozen: packaged--wholesale
5142 Fish, frozen: packaged--wholesale
5146 Seafoods, not canned or frozen packaged--wholesale
5146 Fish, fresh--wholesale
5146 Fish, frozen: except packaged--wholesale
2092 Chowders, fish and seafood: frozen--mfg
2092 Soups, fish and seafood: frozen--mfg
2092 Stews, fish and seafood: frozen--mfg

11 Cadmium
3339 Cadmium refining, primary--mfg
5051 Metals, except precious--wholesale
5051 Nonferrous metal, except precious: e.g, sheets, bars, rods--wholesale
3339 Refining of nonferrous metal, primary: except copper and aluminum--mfg

12 Canned legumes
2032 Baked beans without meat: canned--mfg
2032 Beans, baked: with or without meat--canned--mfg
2033 Canned fruits and vegetables--mfg
2033 Vegetables, canned--mfg
2032 Pork and beans, canned--mfg

13 Computerized Laser Cutting System

3699 Laser welding, drilling and cutting equipment--mfg
3541 Cutting machines, pipe (machine tools)-mfg
3549 Cutting up lines--mfg

14 Electronic thermometers
3823 Thermometers, filled system: industrial process type--mfg
3823 Resistance thermometers and bulbs, industrial process type--mfg
3829 Clinical thermometers, including digital--mfg
3823 Temperature instruments: industrial process type, except glass and bimetal--mfg

15 Oil cleaner/Degreaser
2842 Degreasing solvent--mfg
2842 Solvents, degreasing--mfg
3559 Degreasing machines, industrial--mfg
3559 Degreasing machines, automotive (garage equipment)--mfg

16 POPCORN MACHINES
3556 Corn popping machines, industrial type--mfg
3589 Corn popping machines, commercial type--mfg
3634 Corn poppers, electric--mfg
3634 Popcorn poppers for home use: electric--mfg

17 RADAR DETECTOR/JAMMER
3812 Radar systems and equipment--mfg
3825 Radar testing instruments, electric--mfg
3829 Radiac equipment (radiation measuring and detecting)--mfg
3829 Radiation measuring and detecting (radiac) equipment--mfg

18 SANITARY HARDWARE, BATHROOM ACCESSORIES
3261 Bathroom accessories, vitreous china and earthenware--mfg
3431 Bathroom fixtures: enameled iron, cast iron, and pressed metal--mfg
3088 Bathroom fixtures, plastics--mfg
3431 Sanitary ware: bathtubs, lavoratories, and sinks--metal--mfg
5074 Sanitary ware, china or enameled iron--wholesale
6074 Metal sanitary ware--wholesale

19 Sports Apparel
2253 Sports shirts--mitse--mfg
2311 Tailored dress and sport coats: men's and boys'--mfg
2321 Sport shirts: men's and boys'--mfpm--mfg
2329 Sports clothing, non tailored: men's and boys'--mfpm--mfg
5136 Sportswear, men's and boys'--wholesale
5137 Sportswear: women's and children's--wholesale
2329 Athletic clothing: men's and boys,--mfpm--mfg
2329 Pants, athletic and gymnasium: men's and boys'--mfpm--mfg
2329 Uniforms, athletic and gymnasium: men's and boys'--mfpm--mfg
2339 Athletic uniforms: women's, misses, and juniors'--mfg
2339 Uniforms, athletic: women's, misses', and juniors'--mfpm--mfg
3149 Athletic shoes, except rubber--mfg
5139 Athletic footwear--wholesale
2252 Athletic socks--mfg
2329 Jackets, sport, nontailored: men's and boys'--mfpm--mfg
2329 Riding clothes: men's and boys'--mfpm--mfg

20 Steel Billets
3312 Billets, steel--mfg
5051 Steel--wholesale

21 TEFLON CAR POLISH
7542 Automotive washing and polishing
7542 Washing and polishing, automotive
7542 Waxing and polishing, automotive
2842 Polishes: furniture, automobile, metal, shoe, and stove--mfg
5169 Polishes: furniture, automobile, metal, shoe, etc.--wholesale
2842 Automobile polishes--mfg

22 Wire rods and steel coils
5051 Wire rods--wholesale
5051 Rods, metal--wholesale
3312 Rods, iron and steel: made in steel work or rolling mills--mfg
3312 Steel works producing bars, rods, plates, sheets, structural shapes, etc.--mfg
3312 Nut rods, iron and steel: made in steelworks or rolling mills--mfg

23 Circular Knitting Machinery - bodylength
3552 Knitting machines--mfg
3552 Textile machinery, except sewing machines--mfg

24 Alarm Clocks
5094 Clocks--wholesale
5944 Clocks, including custom made--retail
3873 Clocks, except time clocks--mfg

25 CASHEW NUTS
2068 Nuts, dehydrated or dried--mfg
2068 Nuts: salted, roasted, cooked, or canned--mfg
5145 Nuts, salted or roasted--wholesale
5145 Salted nuts--wholesale
5159 Nuts, unprocessed or shelled only--wholesale

26 CHILDREN'S CRICKET SET
3949 Cricket equipment--mfg
3949 Bats, game: e.g., baseball, softball, cricket--mfg
3949 Sporting goods: except clothing, footwear, small arms, and ammunition--mfg

27 Golf balls
3949 Balls: baseball, basketball, football, golf, tennis, pool, and bowling--mfg
3949 Golfing equipment: e.g., caddy carts and bags, clubs, tees, balls--mfg
5091 Golf equipment--wholesale
5941 Golf goods and equipment--retail

28 Guitar Hangers
3861 Hangers: photographic film, plate, and paper--mfg
3931 Guitars and parts, electric and nonelectric--mfg
5063 Hanging and fastening devices, electrical--wholesale

29 High speed patrol boat
3731 Patrol boats, building and repairing--mfg

30 Lobsters,Live
2092 Seafoods, fresh and frozen--mfg
5146 Seafoods, not canned or frozen packaged--wholesale
0913 Lobsters, catching of

31 Locks and Padlocks
3429 Padlocks--mfg
5072 Locks and related materials--wholesale
5251 Door locks and lock sets--retail
3429 Door locks and lock sets--mfg

32 Salmon in Cans
2091 Salmon: smoked, salted, dried, canned, and pickled--mfg
2091 Fish, canned and cured--mfg
2091 Seafood products, canned and cured--mfg
2091 Canned fish, crustacea, and mollusks--mfg

33 Chocolate and Candy Products
5149 Chocolate--wholesale
2066 Chocolate, sweetened or unsweetened--mfg
2064 Bars, candy: including chocolate covered bars--mfg
2064 Candy bars, except solid chocolate--mfg
2064 Candy, except solid chocolate--mfg
2064 Chocolate bars, from purchased cocoa or chocolate--mfg
2066 Chocolate bars, solid: from cacao beans--mfg
2064 Chocolate candy, except solid chocolate--mfg
2066 Bars, candy: solid chocolate--mfg
2066 Cacao bean products: chocolate, cocoa butter, and cocoa--mfg
2066 Candy, solid chocolate--mfg
5145 Candy--wholesale

34 Concentrated fruits
2087 Concentrates, drink: except frozen fruit--mfg
2087 Fruit juices, concentrated: for fountain use--mfg
2037 Concentrates, frozen fruit juice--mfg
2037 Fruit juice concentrates, frozen--mfg

35 Diluents/Thinners
2851 Lacquer thinner--mfg
2851 Thinner, lacquer--mfg
2851 Thinners, paint: prepared--mfg
2869 Solvents, organic--mfg
2842 Degreasing solvent--mfg
2842 Solvents, degreasing--mfg
2899 Solvents, carbon--mfg

36 SANDALS and SLIPPERS
3021 Shower sandals or slippers, rubber--mfg
3142 House slippers--mfg
3142 Slippers, house--mfg
3149 Sandals, children's: except rubber--mfg
3021 Beach sandals, rubber--mfg
3021 Sandals, rubber--mfg
3149 Footwear, children's: house slippers and vulcanized rubber footwear--mfg

37 Hand Blown Glass
3229 Industrial glassware and glass products, pressed or blown--mfg
3229 Lighting glassware, pressed or blown--mfg
3229 Technical glassware and glass products, pressed or blown--mfg
3229 Scientific glassware, pressed or blown: made in glassmaking plants--mfg

38 SPANISH TOMATO PASTE
2033 Pastes, fruit and vegetable--mfg
2033 Tomato paste--mfg
2033 Sauces, tomato based--mfg

39 CERAMIC TABLEWARE
3229 Tableware, glass and glass ceramic--mfg
3262 China tableware, commercial, and household: vitreous--mfg
3262 Commercial and household tableware and kitchenware: vitreous china--mfg
3262 Hotel tableware and kitchen articles, vitreous china--mfg
3262 Household tableware and kitchen articles, vitreous china--mfg
3262 Tableware, commercial: vitreous china--mfg

40 Canned Food
5149 Canned goods: fruits, vegetables, fish, seafood, meats, and milk--wholesale
2033 Canned fruits and vegetables--mfg
2033 Juices, fruit and vegetable: canned or fresh--mfg
2011 Canned meats, except baby foods and animal feeds--mitse--mfg
2013 Canned meats, except baby foods and animal feeds--mfpm--mfg
2032 Food specialties, canned--mfg
5149 Canned specialties--wholesale
2032 Baby foods (including meats), canned--mfg

41 Golf Shirts
2253 Sports shirts--mitse--mfg
2321 Sport shirts: men's and boys'--mfpm--mfg
5941 Golf goods and equipment--retail

42 Latex examination gloves
3069 Work gloves and mittens: rubber--mfg
2259 Gloves--mitse--mfg
2259 Work gloves and mittens--mitse--mfg
3069 Gloves: e.g., surgeons', electricians', household--rubber--mfg
3842 Gloves, safety: all material--mfg
3842 Safety gloves, all materials--mfg

43 Medical instruments
5047 Medical equipment--wholesale
5047 Surgical and medical instruments--wholesale
3845 Ultrasonic scanning devices, medical--mfg
5047 X-ray machines and parts, medical-- wholesale
3845 Ultrasonic medical equipment, except cleaning--mfg
5047 Diagnostic equipment, medical--wholesale

44 PLASTIC BUCKETS
3089 Buckets, plastics--mfg
3089 Pails, plastics--mfg

45 Ceramic wall and floor tiles
5032 Ceramic wall and floor tile--wholesale
3253 Ceramic tile, floor and wall--mfg
3253 Floor tile, ceramic--mfg
3253 Tile, ceramic wall and floor--mfg
3253 Wall tile, ceramic--mfg
3253 Mosaic tile, ceramic--mfg

46 Roof tiles
3272 Roofing tile and slabs, concrete--mfg
3259 Roofing tile, clay--mfg
3259 Tile, roofing and drain: clay--mfg

47 Magic mop
2392 Mops, floor and dust--mfg
5199 Broom, mop, and paint handles--wholesale
2392 Dust mops--mfg
2392 Floor mops--mfg
7218 Treated mats, rugs, mops, dust tool covers, and cloth supply service

48 Food Dehydrator
3556 Dehydrating equipment, food processing--mfg

49 Syringes
3841 Hypodermic needles and syringes--mfg
3841 Syringes, hypodermic--mfg

50 WOOD PULP
5099 Pulpwood--wholesale
2611 Fiber pulp: made from wood, rags, wastepaper, linters, straw, and bagasse--mfg
2611 Pulp, fiber: made from wood, rags, wastepaper, linters, straw, and bagasse--mfg
2611 Wood pulp--mfg

51 GLASS BOWL
3229 Bowls, glass--mfg
3229 Tableware, glass and glass ceramic--mfg

52 Catalytic Converters
7533 Catalytic converters, automotive: installation, repair, or sales and installation

53 ORGANIC GRANOLA CEREAL
2043 Granola, except bars and clusters--mfg
2043 Cereal preparations and breakfast foods--mfg
2043 Breakfast foods, cereal--mfg

54 Pressure cooker
3469 Pressure cookers, stamped or drawn--mfg
3589 Pressure cookers, steam: commercial--mfg
3365 Pressure cookers, domestic: cast aluminum, except die castings--mfg
3469 Cookers, pressure: stamped or drawn--mfg

55 Shower Curtain
2392 Curtains, shower--mfpm--mfg

2392 Shower curtains--mfpm--mfg

56 Military Blankets
2392 Blankets--mfpm--mfg
5023 Blankets--wholesale
2211 Blankets and blanketings, cotton--mitse--mfg
2231 Blankets and blanketings, wool and similar animal fibers--mitse--mfg
2311 Military uniforms, tailored: men's and boys'--mfg

57 Bagel Chips
2051 Bagels--mfg
2096 Corn chips and related corn snacks--mfg
2096 Potato chips and related corn snacks--mfg
5145 Corn chips--wholesale
5145 Potato chips--wholesale

58 Condensed Milk
2023 Condensed and evaporated milk--mfg
2023 Milk: concentrated, condensed, dried, evaporated, and powdered--mfg
2023 Buttermilk: concentrated, condensed, dried, evaporated, and powdered--mfg
2023 Evaporated milk--mfg

59 Jet fuel
2911 Jet fuels--mfg
2911 Fuels, jet--mfg

60 Video/Audio Cassettes
5065 Cassettes, recording--wholesale
5099 Cassettes, prerecorded: audio--wholesale
3695 Video recording tape, blank--mfg
5065 Tapes, audio and video recording-wholesal
7822 Tapes, video, recorded--wholesale
7822 Video tapes, recorded--wholesale
3695 Audio range tapes, blank--mfg
5099 Tapes, audio prerecorded--wholesale

61 Polypropylene and Jute Bags
2299 Bagging, jute: made in jute weaving mills--mfg
2392 Bags, blanket: plastics--mfg
2392 Blanket bags, plastic--mfg
2673 Merchandise bags, plastics--mfpm--mfg
2759 Bags, plastics: printed only, except lithographed or gravure (bags not made in printing plants)--mfg
5113 Bags, paper and disposable plastics--wholesale

62 Powdered Milk Blend
2023 Ice milk mix, unfrozen: liquid or dry--mfg
2023 Ice cream mix, unfrozen: liquid or dry--mfg
2023 Dry milk products: whole milk, nonfat milk, buttermilk, whey, and cream--mfg
2023 Yogurt mix--mfg

63 Metal ingot
3339 Ingots, primary: nonferrous metals, except copper and aluminum--mfg
3341 Ingots, nonferrous: smelting and refining--secondary--mfg
5051 Ingots--wholesal

3339 Pigs, primary: nonferrous metals, except copper and aluminum--mfg

64 Organic Chemicals
5169 Organic chemicals, synthetic--wholesale
2869 Laboratory chemicals, organic--mfg
2869 High purity grade chemicals, organic: refined from technical grades--mfg
2869 Organic chemicals, acyclic--mfg
2869 Reagent grade chemicals, organic: refined from technical grades, except diagnostic and substances--mfg
2833 Chemicals, medicinal: organic and inorganic--bulk, uncompounded--mfg
2833 Organic medicinal chemicals: bulk--mfg

65 WOODEN TOYS and MARIONETTES
3999 Marionettes (puppets)--mfg
3999 Puppets--mfg
3944 Blocks, toy--mfg

66 animals - stuffed
3942 Stuffed toys (including animals)--mfg
3942 Toys, stuffed--mfg

67 gold or platinum ingot watches
3915 Watch jewels--mfg
5094 Watches and parts--wholesale
5944 Watches, including custom made--retail

68 electronic connectors
3678 Connectors, electronic: e.g., coaxial, cylindrical, rack and panel, printed circuit--mfg
5065 Electronic connectors--wholesale
5065 Connectors, electronic--wholesale

69 Siphon Pump
5084 Pumps and pumping equipment, industrial--wholesale
3561 Pumps, general industrial type--mfg
3594 Pumps for fluid power systems--mfg
3594 Pumps, hydraulic power transfer--mfg
3561 Domestic water pumps--mfg

70 Baseball cards
2752 Souvenir cards, lithographed--mfg
2759 Souvenir cards: except lithographed or gravure--mfg
2754 Souvenir cards: gravure printing--mfg

71 Aluminum Gas Cylinders
3443 Cylinders, pressure: metal plate--mfg
3593 Pneumatic cylinders, fluid power--mfg
3593 Cylinders, fluid power: hydraulic and pneumatic--mfg
3593 Fluid power cylinders, hydraulic and pneumatic--mfg

72 Baby dresses
2361 Dresses: girls', children's, and infants'--mfpm--mfg
5137 Clothing: women's, children's, and infants'--wholesale

73 Honey Powder

2087 Flavoring extracts, pastes, powders, and syrups--mfg
5149 Honey--wholesale

74 Instant Pregnancy Tests
2835 Pregnancy test kits--mfg

75 LAUNDRY BALLS - SOAP SUBSTITUTE
5169 Laundry soap, chips, and powder-wholesale
5169 Soap, chips, and powder: laundry-wholesale
2841 Soap: granulated, liquid, cake, flaked, and chip--mfg

76 Oil spills Cleanup - Oil Terminator
4959 Oil spill cleanup

77 Rubber Synthetic Isoprene
2822 Isoprene rubbers, synthetic--mfg
2822 Isobutylene-isoprene rubbers--mfg
5169 Synthetic rubber--wholesale

78 Synthetic Stucco
5032 Stucco--wholesale
3299 Stucco--mfg
5032 Plaster--wholesale

79 Tank Trucks
3795 Tank recovery vehicles--mfg
3743 Tank freight cars and car equipment--mfg

80 olive oil
2079 Oil, olive--mfg
2079 Olive oil--mfg
2076 Oils, vegetable: except corn, cottonseed, and soybean--mfg
2079 Cooking oils, vegetable: except corn oil--refined--mfg
2079 Oils, vegetable (except corn oil) refined: cooking and salad--mfg
2079 Vegetable cooking and salad oils, except corn oil: refined--mfg
5149 Vegetable cooking oil--wholesale
2833 Vegetable oils, medicinal grade: refined and concentrated--mfg

81 Turbine Generator
3511 Turbo-generators--mfg
3511 Gas turbine generator set units, complete--mfg
3511 Generator set units, turbine: complete--steam, gas, and hydraulic--mfg
3511 Hydraulic turbine generator set units, complete--mfg
3511 Solar powered turbine-generator sets--mfg
3511 Steam turbine generator set units, complete--mfg
3511 Turbine generator set units, complete: steam, gas, and hydraulic--mfg
3511 Wind powered turbine-generator sets--mfg

82 AUTOMATIC WRIST BLOOD PRESSURE MONITOR
3841 Blood pressure apparatus--mfg
7299 Blood pressure testing, coin operated

83 Assorted Diaries and Desk Calendars
2782 Diaries--mfg

3999 Calendars, framed--mfg
2741 Calendars: publishing and printing, or publishing only--mfg
2759 Calendars, printed: except lithographed or gravure--mfg

84 Auto Batteries
5013 Batteries, automotive--wholesale

85 Fragrances-Incense Sticks and Sprays
2899 Incense--mfg
5122 Perfumes--wholesale
2844 Perfumes, natural and synthetic--mfg
2869 Perfume materials, synthetic--mfg

86 Frozen Turkeys
2015 Turkeys, processed: fresh, frozen, canned, or cooked--mfg
5142 Poultry, frozen: packaged--wholesale
5144 Poultry: live, dressed, or frozen (except packaged)--wholesale
2015 Poultry, processed: fresh, frozen, canned, or cooked--mfg
2015 Chickens, processed: fresh, frozen, canned, or cooked--mfg
2015 Ducks, processed: fresh, frozen, canned, or cooked--mfg

87 Hair Accessories
5131 Hair accessories--wholesale
3999 Hair goods: braids, nets, switches, toupees, and wigs--mfg
3634 Driers: hand, face, and hair--electric--mfg
3634 Hair dryers, electric: except equipment designed for beauty parlor use--mfg
3999 Clippers, hair: for human use--hand and electric--mfg
3999 Hair clippers for human use, hand and electric--mfg
3965 Hairpins, except rubber--mfg

88 Lentils and Chickpeas
5148 Vegetables, fresh--wholesale
0119 Lentil farms
0161 Vegetable farms
0161 Pea farms, except dry peas

89 Apple Juice
2033 Fruit juices: canned--mfg
2033 Juice, fruit: concentrated-hot pack--mfg
2033 Juices, fresh: fruit or vegetable--mfg
2033 Juices, fruit and vegetable: canned or fresh--mfg

90 COTTON FLANNEL
2211 Canyon flannels, cotton--mfg
2211 Flannels, cotton--mfg
2211 Mitten flannel, cotton--mfg
2211 Outing flannel, cotton--mfg
2211 Broadwoven fabrics, cotton--mfg
2211 Gabardine, cotton--mfg

91 BALLPOINT PENS
3951 Ball-point pens--mfg
3951 Pens and pen parts: fountain, stylographic, and ball-point--mfg

92 Calcium Carbide
2819 Calcium carbide, chloride, and hypochlorite--mfg
2819 Calcium compounds, inorganic--mfg
2819 Carbide--mfg

93 Stainless steel dishes
3914 Hollowware: silver, nickel silver, pewter, stainless steel, and plated--mfg
3914 Trays: silver, nickel silver, pewter, stainless steel, and plated--mfg

94 sesame seeds
0181 Seeds, flower and vegetable: growing of
5191 Seeds: field, garden, and flower--wholesale

95 air compressors
3563 Compressors, air and gas: for general industrial use--mfg
5084 Compressors, except air-conditioning and refrigeration--wholesale

References

Agirre, E. and Rigau, G., 1995, A proposal for Word Sense Disambiguation Using Conceptual Distance, *Proceedings of the First International Conference on Recent Advanced in NLP,* Bulgaria.

Anderberg, M.R., 1973, *Cluster Analysis for Applications*, Academic Press, New York.

Austin, J.L., 1961, Philosophical Papers, Oxford University Press.

Barsalou, L.W., and Hale, C.R., 1993, Components of Conceptual Representation: from Feature Lists to Recursive Frames, in *Categories and Concepts: Theoretical Views and Inductive Data Analysis,* Edited by I.V. Mechelen, J. Hampton, R.S. Michalski, and P. Theuns, Academic Press Ltd. 97-144.

Batagelj, V. and Bren, M., 1995, Comparing Resemblance Measures, *Journal of Classification,* Vol. 12, 73-90.

Blosseville, M.J., Hebrail, G., Monteil, M.G., and Penot, N., 1992, Automatic Document Classification: Natural Language Processing, Statistical Analysis, and Expert System Techniques Used Together, *Proceedings of the 15th Annual International Conference on Research and Development in Information Retrieval,* SIGIR'92.

Bookman, L. A., 1994, *Trajectories through Knowledge Space: A Dynamic Framework for Machine Comprehension,* Boston: Kluwer Academic Publishers.

Bruce, R. and Wiebe, J., 1994, Word-Sense Disambiguation Using Decomposable Models, *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics,* 139-146.

Chakravarthy, A.S. and Haase, K. B., 1995, NetSerf: Using Semantic Knowledge to Find Internet Information Archives, *Proceedings of the 18th Annual International Conference on Research and Development in Information Retrieval,* SIGIR'95, 4-11.

Charniak, E., 1993, *Statistical Language Learning*. MIT Press.

Chater, N. and Hahn, U., 1997, Representational Distortion, Similarity and the Universal Law of Generalization, *Proceedings of SimCat'97.*

Chen, H., Dhar, V., 1990, A Knowledge-Based Approach to the Design of Document-Based Retrieval Systems, *Proceedings of Conference on Office Information Systems*, April 1990, ACM SIGOIS Bulletin, Vol. 11, No. 2&3, 281-290.

Chen, H., Lynch, K. J., Basu, K., and Ng, T.D. 1993, Generating, Integrating, and Activating Thesauri for Concept-Based document Retrieval, *IEEE Expert*, Vol. 18, No. 2, 25-34.

Church, K.W. and Hanks, P., 1989, Word Association Norms, Mutual Information, and Lexicography, *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, ACL27'89, 76-83.

Conrad, J. G. and Utt, M. H., 1994, A System for Discovering Relationships by Feature Extraction from Text Databases, *Proceedings of the 17th Annual International Conference on Research and Development in Information Retrieval*, SIGIR'94.

Conrath, D., 1993, The MarketPlace: Concepts and Issues, *Journal of Small Business and Entrepreneurship*, Vol. 10, No. 4, 69-80.

Conrath, D., 1994, The MarketPlace: a Computer-based, Global Trading network, *Journal of Enterprising Culture*, Vol. 1, No. 3&4, 383-402.

Cook D.J. and Holder, L.B. 1994, Substructure discovery using minimum description length and background knowledge, *Journal of Artificial Intelligent Research*, Vol.1.

Corter, J.E., 1996, *Tree Models of Similarity and Association*, Sage University Paper.

Dagan I, and Itai, A., 1994, Word Sense Disambiguation Using a Second-Language Monolingual Corpus, *Computational Linguistics*, Vol. 20, No.4, 563-596.

Dagan I., Marcus S. and Markovitch S., 1993, Contextual Word Similarity and Estimation from Sparse Data, *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, ACL31'93.

Dagan, I., Pereira, F., and Lee, L. 1994, Similarity-based Estimation of Word Cooccurrence Probabilities, *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, ACL32'94.

Dagan, I., Lee, L., and Pereira, F, 1997, Similarity-based Methods for Word Sense Disambiguation,, *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, ACL97.

Delugach, H.S., 1992, An Exploration into Semantic Distance, *Proceedings of the 7th Annual workshop on Conceptual Structure: Theory and Implementation*, Las Cruces, MN, July 1992, 119-124.

Desmarais, N., 1992, Gold Turns to Silver: Yellow Pages on CD-ROM, *CD-ROM Librarian*, Vol. 7, No. 3, 21-27.

Dice, L. R., 1945, Measures of the amount of ecological association between species, *Ecology*, Vol. 26, 297-302.

Downie, N. M. And Starry, A. R., 1977, *Descriptive and Inferential Statistics*, Harper & Row.

Downing, P., 1977, On the Creation and the use of English Compound Nouns, *Language*, Vol. 53, No. 4, 810-842.

Driscoli, J., Lautenschlager, J. and Zhao, M., 1993, The QA system, *Proceedings of TREC-1*. 199-207.

Dunning, T., 1993, Accurate Methods for the Statistics of Surprise and Coincidence, *Computational Linguistics*, Vol.19, No. 1,61-74.

Estes, W.K., 1994, *Classification and Cognition*, Oxford University Press.

Fano, R. 1961, *Transmission of Information*, MIT Press, Cambridge, Massachusetts.

Feger, H. and De Boeck P, 1993, Categories and Concepts: Introduction to Data Analysis, in *Categories and Concepts: Theoretical Views and Inductive Data Analysis*, Edited by I.V. Mechelen, J. Hampton, R.S. Michalski, and P. Theuns, Academic Press Ltd. 204-223.

Fellbaum, C., 1998, *WordNet: An Electronic Lexical Database and Some of its Applications*, The MIT Press.

Gale, W., Church K., and Yarowsky, D. 1992, A Method for Disambiguating Word Senses in a Large Corpus, *Computers and the Humanities*, Vol. 26, 415-439.

Gay, L.S., and Croft, W.B., 1990, Interpreting Nominal Compounds for Information Retrieval, *Information Processing & Management*, Vol. 26, No.1, 21-38, 1990.

Gentner, D., 1983, Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, Vol. 7, 155-170.

Gentner, D. and Markman, A.B., 1995, Similarity is like Analogy: Structural Alignment in Comparison, *Similarity in Language, Thought and Perception*, Edited by C. Cacciari, Brepols.

Gentner, D. and Markman, A.B., 1997, Structure-mapping in analogy and similarity, *American Psychologist*, Vol. 52, No.1, 45-56.

Ginsberg, A., 1993, A Unified Approach to Automatic Indexing and Information Retrieval, *IEEE Expert*, Vol. 8, No. 5, 46-56.

Girardi, M.R. and B. Ibrahim, 1994, A Similarity Measure for Retrieving Software Artifacts, *Proceedings of Sixth International Conference on Software Engineering and Knowledge Engineering (SEKE'94)*, Latvia, June 21-23, 1994, 478-485.

Glucksberg, S and Manfredi D., 1995, Metaphoric Comparisons, *Similarity in Language, Thought and Perception*, Edited by C. Cacciari, Brepols.

Goertzel, B., 1997, Similarity as Transformation, in *Mind as a Complex System*, Ed. Goertzel et al., http://www.goertzel.org/ben/catpap.html.

Goertzel B. and Kalish, M., 1996, *The Attack of the Aliens from Vector Space: Steps Toward a Complex Systems Theory of Categorization and Similarity*, http://www.goertzel.org/ben/catpap.html

Goldstone, R. L., 1994, Similarity, Interactive Activation, and Mapping, *Journal of Experimental Psychology: Learning, Memory, and Cognition*, Vol. 20, 3-28.

Goldstone, R. L., 1996, Alignment-based nonmonotonicities in similarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, Vol. 22, 988-1001.

Goldstone, R. L., 1998, Similarity. in R.A. Wilson & F. C. Keil (eds.) *MIT Encyclopedia of the Cognitive Sciences*. Cambridge, MA: MIT Press, online version: http://mitpress.mit.edu/MITECS/.

Goldstone, R. L., in press, Hanging Together: A connectionist model of similarity. In J. Grainger, A. Jacobs, & J. Townsend (Eds.) *Symbolic Connectionism*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Goldstone, R. L., Medin, D. L., and Gentner, D., 1991, Relational similarity and the nonindependence of features in similarity judgments, *Cognitive Psychology*, Vol. 23, 222-264.

Gower, J. C., 1971, A General Coefficient of Similarity and Some of its Properties, *Biometrics*, Vol. 27, 857-874.

Gower, J. C., 1985, Measures of Similarity, Dissimilarity, and Distance, *Encyclopaedia of Statistical Science*, Vol.5, 397-405.

Gower, J.C. and Legendre, P., 1986, Metric and Euclidean Properties of Dissimilarity Coefficients, Journal of Classification, Vol. 3, 5-48.

Grefenstette, G., 1992, Use of Syntactic Context to Produce Term Association Lists for Text Retrieval, *Proceedings of the 15th Annual International Conference on Research and Development in Information Retrieval*, SIGIR'92.

Grefenstette, G., 1994, *Explorations in Automatic Thesaurus Discovery*, Kluwer Academic Publishers.

Gregory J. C., 1987, *Algorithmic Information Theory*, Cambridge University Press.

Gregson, R., 1975, *Psychometrics of Similarity*, Academic Press, New York.

Hampton, J. 1993, Prototype Models of Concept Representation, in *Categories and Concepts: Theoretical Views and Inductive Data Analysis*, Edited by I.V. Mechelen, J. Hampton, R.S. Michalski, and P. Theuns, Academic Press Ltd. 67-95.

Hahn, U. and Chater, N., 1997, Concepts and Similarity. In K. Lamberts and D. Shanks (Eds.), *Knowledge, Concepts and Categories*. London: UCL Press

Harman, D., 1995, Overview of the Third Text Retrieval Conference. *Proceedings of the Third Text Retrieval Conference (TREC-3).*

Hayes, P., 1992, Intelligent High-Volume Text Processing Using Shallow, Domain-Specific Techniques, *Text-Based Intelligent Systems*, Edited by Paul Jacobs, 227-241.

Hindle, D., 1990, Noun Classification from Predicate-Argument Structures, *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, ACL28'90, 268-275.

Hindle, D and Rooth, M., 1993, Structural Ambiguity and Lexical Relations. *Computational Linguistics*, Vol.19, 103-121.

Hirst, G., 1995, Near-synonymy and the Structure of Lexical Knowledge, *Proceedings of the AAAI Symposium on Representation and Acquisition of Lexical Knowledge: Polysemy, Ambiguity, and Generativity*, 51-56, Stanford, CA.

Hirst, G. and St-Onge, D., 1998, Lexical Chains as Representations of Context for the Detection and Correction of Malapropism, *in WordNet: An Electronic Lexical Database*, edited by Christiane Fellbaum, The MIT Press.

Holyoak, K. J., and Thagard, P., 1989, Analogical mapping by constraint satisfaction. Cognitive Science, Vol. 13, 295-355.

Hoppe, H. U., Ammersbach, K. Lutes-Schaab, B., and Zinbmeister, G. 1990, EXPRESS: An Experimental Interface for Tactual Information Retrieval, *Proceedings of the 13th Annual International Conference on Research and Development in Information Retrieval*, SIGIR'90. 63-81, Brussels.

Jaccard, P., 1908, Nouvelles recherches sur la distribution florale, *Bulletin de la Societe vaudoise des Sciences Naturelles*, VoL 44, 223-270.

Jiang, J., 1996, *Semantic Association Based on Corpus Analysis and Lexical Taxonomy*, Technical Report, 204-MS-1996, Department of Management Sciences, University of Waterloo, Canada.

Jiang, J. and Conrath, D., 1996, A Concept-based Approach to Retrieval from an Electronic Industrial Directory, *International Journal of Electronic Commerce*, VoL 1, No.1, 51-72

Jiang, J. and Conrath, D., 1997, Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy, *Proceedings of International Conference Research on Computational Linguistics (ROCLING X)*, 19-33.

Jing, Y., and Croft, W.B., 1994, *An Association Thesaurus for Information Retrieval*, Technical Report, IR-47, Department of Computer Sciences, University of Massachusetts.

Johansson, C., 1994, Catching the Cheshire Cat, *Proceedings of the 17th International Conference on computational Linguistics*, COLING'94.

Jones, K.P., 1989, The Notion of a Minimum Unit for Information Retrieval Systems, in *Proceedings of Informatics 10: Prospects for Intelligent Retrieval*, Cambridge, Aslib, 91-105.

Jones, S., 1993, A Thesaurus Data Model for an Intelligent Retrieval System, *Journal of Information Science*, Vol. 19, 167-178.

Jurisica, I., 1994, *Context-Based Similarity Applied to Retrieval of Relevant Cases*, Technical Report, CBR-TR94-5, University of Toronto, Canada.

Kaufman, L. and Rouss, P.J., 1990, *Finding Groups in Data : an Introduction to Cluster Analysis*, John Wiley & Sons Inc.

Khoo, C.S.G. and Poo, D.C.C., 1994, An Expert System Approach to Online Catalog Subject Searching, *Information Processing & Management*, VoL 30, No. 2, 223-238.

Kimoto, H. and Iwadera, T., 1990, Construction of a Dynamic Thesaurus and its Use for Associated Information Retrieval, *Proceedings of the 13th Annual International Conference on Research and Development in Information Retrieval*, SIGIR'90, 227-240.

Kozima, H. and Furugori, T., 1993, Similarity between words computed by spreading activations on an English dictionary, *Proceedings of the 5th Conference of the European Chapter of the Association for Computational Linguistics*, EACL-93, 232-239.

Kresge, D. T., 1990, Expansion of SIC Codes for More Precise Searching, *Proceedings of the 11th National Online Meeting*, Ed. Martha E. Williams, (May 1990), New York: Learned Information, Inc.

Kristensen, J., 1993, Expanding End-user's Query Statements for Free Text Searching with a Search-aid Thesaurus, *Information Processing & Management*, Vol. 29, No. 6, 733-744.

Krovetz, R., 1990, Viewing the Dictionary as a Classification System, *Proceedings of the First ASIS Classification Workshop*, 87-93.

Lakoff, G., 1987, *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*, University of Chicago Press.

Lauer, M., 1994, Conceptual Association for Compound Noun Analysis, *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, ACL'94, 337-339.

Lauer, M., 1995a, Corpus Statistics Meet the Noun Compound: Some Empirical Results, *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, ACL'95, Boston.

Lauer, M. 1995b, *Designing Statistical Language Learners: Experiments on Noun Compounds*, Ph.D. dissertation, Macquarie University, Australia.

Lauzon, D. and Rose, T., 1994, Task-Oriented and Similarity-Based Retrieval, *Proceedings of the Ninth Knowledge-based Software Engineering Conference*, 98-107.

LDOCE, 1987, *Longman Dictionary of Contemporary English*, Longman, Harlow, Essex, new edition.

Leacock, C., Towell, G., and Voorhees, E., 1993, Corpus-based Statistical Sense Resolution, *Proceedings of ARPA Workshop on Human Language Technology*, 260-265, March 1993.

Ledwith, R. H., 1988, Development of a Large, Concept-Oriented Database for Information retrieval, *Proceedings of the 11th Annual International Conference on Research and Development in Information Retrieval*, SIGIR'88, 651-661.

Lee, J.H., Kim, M.H., and Lee, Y. J., 1993, Information Retrieval Based on Conceptual Distance in IS-A Hierarchies, *Journal of Documentation*, Vol. 49, No. 2, 188-207.

Leonard, R., 1984, *The Interpretation of English Noun Sequences on the Computer*, North-Holland, Amsterdam.

Lesk, M., 1986, Automatic Sense Disambiguation: How to Tell a Pine Cone from an Ice Cream Cone, *Proceedings of the SIGDOC Conference*, 24-26.

Levi, J. 1978, *The Syntax and Semantics of Complex Nominals*, Academic Press, New York.

Liddy, E. D., and Myaeng, S. H., 1993, DR-LINK's: Linguistic-Conceptual Approach to Document Detection, *Proceedings of TREC-1*, 1993, 113-129.

Liddy, E. D., and Myaeng, S. H., 1994, DR-LINK: A System Update for TREC-2, *Proceedings of TREC-2*, 1994, 85-99.

Lin, D., 1997, Using syntactic dependency as local context to resolve word sense ambiguity, *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, ACL97.

Luk, A. K., 1995, Statistical sense disambiguation with relatively small corpora using dictionary definitions, *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, ACL'95.

Malone, Thomas W.; Yates, Joanne; and Benjamin, Robert I. ,1987, Electronic markets and Electronic Hierarchies. *Communication of ACM*, Vol. 30, No. 6, 484-497.

Markman, A. B. and Gentner, D. 1993, Structural alignment during similarity comparisons, *Cognitive Psychology*, Vol. 25, 431-467.

Mauldin, M. L., 1991, *Conceptual Information Retrieval: A case study in Adaptive Partial Parsing*, Kluwer Academic Publishers, Boston.

McCarthy, D., 1997, Word sense disambiguation for acquisition of selectional preferences, Proceedings of ACL/EACLWorkshop.

McDonald, D.B. 1982, *Understanding Noun Compounds*, PhD thesis, Carnegie Mellon University, Pittsburgh, PA.

McKeown, K. 1993, Augmenting lexicons automatically: clustering semantically related adjectives, *Proceedings of ARPA Workshop on Human Language Technology*, 272-277, March 1993.

McMath, C. F., Tamaro, R.S., and Rada R., 1989, A Graphical Thesaurus-based Information Retrieval, *International Journal of Man-Machine Studies*, Vol. 31, 121-147.

Medin, D. L., Goldstone, R. L., and Gentner, D., 1990, Similarity involving attributes and relations: Judgements of similarity and difference are not inverses. *Psychological Science*, Vol.1, No.1, 64-69.

Medin, D. L., Goldstone, R. L., and Gentner, D., 1993, Respects for Similarity, *Psychological Review*, Vol. 100, No. 2, 254-278.

Medin, D. L., Goldstone, R. L., 1995, The Predicates of similarity, in *Similarity in Language, Thought and Perception*, Edited by C. Cacciari, Brepols.

Miller, G., 1990, Nouns in WordNet: A lexical inheritance system, *International Journal of Lexicography*, Vol. 3, No. 4, 245-264.

Miller, G. and Charles, W.G., 1991, Contextual correlates of semantic similarity, *Language and Cognitive Processes*, Vol. 6, No. 1, 1-28.

Miller, G., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K., 1990, Introduction to WordNet: An online lexical database, *International Journal of Lexicography*, Vol. 3, No. 4, 235-244.

Miller, G., Leacock, C, Tengi, R., and Bunker, R.T., 1993, A Semantic Concordance, *Proceedings of ARPA Workshop on Human Language Technology*, 303-308.

Miller, G., Chodorow, M., Landes, S., Leacock, C., and Thomas, R.G., 1994, Using a Semantic Concordance for Sense Identification, *Proceedings of ARPA Workshop on Human Language Technology*, 240-243.

Montgomery, D. C., 1991, *Design and Analysis of Experiments*, third edition, John Wiley & Sons.

Morris, J. and Hirst, G., 1991, Lexical cohesion computed by thesaural relations as an indicator of the structure of text, *Computational Linguistics*, Vol. 17, 21-48.

Murphy, G. L. and Medin, D.L., 1985, The role of Theories in Conceptual Coherence, *Psychological Review*, Vol. 92, 289-316.

Myaeng, S.H. and Li, M., 1992, Building Term Clustering by Acquiring Lexical Semantics from a Corpus, *Proceedings of the First International Conference on Information and Knowledge Management*, CIKM'92, 130-137.

Ng, H. T., 1997a, Getting Serious about Word Sense Disambiguation, , In *Proceedings of SIGLEX '97*, 1-7, Washington, DC,

Ng, H. T., 1997b, Exemplar-Based Word Sense Disambiguation: Some Recent Improvements, *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing (EMNLP-2)*, August 1997

Ng, H. T., and Lee, H. B., 1996, Integrating Multiple Knowledge Sources to Disambiguate Word Sense: An Exemplar-Based Approach, *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, ACL'96.

Niwa, Y. and Nitta, Y. 1994, Co-occurrence vectors from corpora vs. distance vectors from dictionaries, , *Proceedings of the 17th International Conference on computational Linguistics*, COLING'94, 304-309.

Nosofsky, R. M., 1992, Similarity scaling and cognitive process models, *Annual Review of Psychology*, Vol. 43, 25-53.

Noy, N. F. and Hafner, C. D., 1997, The Sate of the Art in Ontology Design: A Survey and Comprehensive Review, *AI Magazine*, Vol. 18, No. 3., 53-74.

Nutter, J.T., 1989, *A Lexical Relation Hierarchy*, Technical Report TR-89-06, Virginia Polytechnic Institute and State University.

Ogden, C.K. and Richards, K., 1923, *The Meaning of Meaning*, London, Routledge & Kegan Paul Ltd.

Old, L. J., 1996, Synonymy and Word Equivalence, *Proceedings of Conference of Midwest Artificial Intelligence and Cognitive Science.*

Paice, C. D., 1991, A Thesaural Model of Information Retrieval, *Information Processing & Management*, Vol. 25, No. 5, 433-447.

Peh, L.S. and Ng, H.T, 1997, Domain-Specific Semantic-Class Disambiguation Using WordNet, *Proceedings of the Fifth Workshop on Very Large Corpora*, 56-64.

Priss, U. E., 1996, Classification of Meronymy by Methods of Relational Concept Analysis, *Proceedings of Midwest Artificial Intelligence and Cognitive Science Conference.*

Pustejovsky, J. 1993, *Semantics and the Lexicon*, Kluwer Academic Publishers.

Quirk, R., Greenbaun, S, Leech, G., and Svartvik, J., 1985, *A Comprehensive Grammar of the English Language*, Longman, New York.

Rada, R., and Bicknell. E., 1989, Ranking Documents with a Thesaurus, *Journal of the American Society for Information Science*, Vol. 40, No. 5, 304-310.

Rada R., Mili, H., Bicknell. E., and Bletner, M., 1989, Development and Application of a Metric on Semantic Nets, *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 19, No. 1, 17-30.

Resnik, P., 1992a, WordNet and Distributional Analysis: A Class-based Approach to Lexical Discovery, *Proceedings of the AAAI Symposium on Probabilistic Approaches to Natural Language*, San Joe, CA.

Resnik P., 1992b, A Class-based Approach to Lexical Discovery, *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, ACL30th'92.

Resnik, P., 1993a, Selection and Information: A Class-Based Approach to Lexical Relationships, Ph.D. Dissertation, University of Pennsylvania.

Resnik, P., 1993b, Semantic Classes and Syntactic Ambiguity, *Proceedings of ARPA Workshop on Human Language Technology*, 278-283, March 1993.

Resnik, P., 1995, Using Information Content to Evaluate Semantic Similarity in a Taxonomy, *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Vol. 1, 448-453, Montreal, August 1995.

Resnik, P., 1997, Selectional Preference and Sense Disambiguation, In *Proceedings of SIGLEX '97*, Washington, DC, pp52-57.

Resnik, P. and D. Yarowsky, 1997 A Perspective on Word Sense Disambiguation Methods and Their Evaluation, In *Proceedings of SIGLEX '97*, Washington, DC, pp. 79-86.

Ribas, F., 1995, On learning more appropriate selectional restrictions, *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics*, Dublin, Ireland.

Richardson, R., and Smeaton, A.F., 1995a, *Using WordNet in a Knowledge-Based Approach to Information Retrieval*, Working Paper, CA-0395, School of Computer Applications, Dublin City University, Ireland.

Richardson, R., and Smeaton, A.F., 1995b, *Automatic Word Sense Disambiguation in a KBIR Application*, Working Paper, CA-0595, School of Computer Applications, Dublin City University, Ireland.

Rigau, G., Atserias, I, and Agirre, E., 1997, Combining Unsupervised Lexical Knowledge Methods for Word Sense Disambiguation, *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics* ACL'97.

Riloff, E., 1993, Automatic Constructing a Dictionary for Information Extraction Tasks, *Proceedings for the 11th National Conference on Artificial Intelligence*, 811-816.

Rips, L.J., 1989, Similarity, Typicality, and Categorization, in *Similarity and Analogical Reasoning,* edited by Stella Vosniadou, Andrew Ortony, Cambridge University Press, 21-59.

Rips, L.J., and Collins, A., 1993, Categories and resemblance. *Journal of Experimental Psychology: General*, Vol.122, 468-486.

Roget's 1977, *Roget's International Thesaurus*, Harper & Row, New York, Fourth Edition.

Rosch, E. and Lloyd, B. B., 1978, *Cognition and Categorization*, Lawrence Erlbaum Associates.

Ross, S., 1976, A First Course in Probability, Macmillan.

Ruge, G., 1992, Experiments on linguistically-based term associations, *Information Processing & Management*, Vol. 28, No. 3, 317-332.

Ruge, G., 1995, Human memory models and term association, *Proceedings of the 18th Annual International Conference on Research and Development in Information Retrieval*, SIGIR'95, 219-227.

Ryder, M., 1994, *Ordered Chaos: The Interpretation of English Noun-Noun Compounds*, Berkeley/Los Angeles/ London: University of California Press.

Salton, G., and McGill, M. J. 1983, *Introduction to Modern Information Retrieval*, McGraw-Hill

Schmits-Esser, W. 1991, New Approaches in Thesaurus Application, *International Classifications*, Vol. 18, No. 3, 143-147.

Sedelow, S.Y. and Sedelow, W.A., 1994, Thesauri and Concept-Lattice Semantic Nets, *Advances in Knowledge Organization*, Vol. 4, 350-357.

Shaw, D., 1991, The Human-Computer interface for Information Retrieval, *Annual Review of Information Science and Technology (ARIST)*, Vol. 26, 155-195.

Shepard, R.N., 1962, The analysis of proximities: multidimensional scaling with an unknown distance function, part I, *Psychometrika*, Vol. 27, 125-140.

Shepard, R.N., 1987, Toward a universal law of generalization for psychological science, *Science*, Vol. 237, 1317-1323.

Silvester, J.P., Genuard, M.T., and Klingbiel, P.H., 1994, Machine-aided Indexing at NASA, *Information Processing & Management*, Vol. 30, No. 5, 631-645.

Sinclair, J. editor, 1987, *Collins COBUILD English Language Dictionary*, Collins and the University of Birmingham, London.

Smadja, F.A. and McKeown, K.R. 1990, Automatically extracting and representing collocations for language generation, *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, ACL'90, 252-259.

Smeaton, A.F., 1989, Information Retrieval and Natural Language Processing, *Informatics 10: Prospects for Intelligent Retrieval*, Ed. Kevin P. Jones, Cambridge, Aslib, 1989, 1-14.

Smeaton, A.F., 1997, Information Retrieval: Still Butting Heads with Natural Language Processing? in: *Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology*, M.T Pazienza (Ed.), Springer-Verlag Lecture Notes in Computer Science, 115-138.

Smeaton, A.F. and Quigley, I., 1996, Experiments on Using Semantic Distance Between Words in Image Caption Retrieval, in: *Proceedings of the 19th International Conference on Research and Development in Information Retrieval (SIGIR96)* Zurich, Switzerland, August 1996, 174-180,.

Sneath, P.H. and Sokal, R.R., 1973, *Numerical Taxonomy: the Principles and Practice of Numerical Classification*, W.H. Freeman and Company

Statistics Canada, 1980, *Standard Industrial Classification Manual*, Statistics Canada, Standard Division, December 1980.

Statistics Canada, 1993, *Standard Classification of Goods*, Statistics Canada, Standard Division, March 1993.

Steier, A.M. and Belew, R.K., 1993, Exporting Phrases: A Statistical Analysis of Topical Languages, *Proceedings of the 2nd Symposium on Document Analysis and Information Retrieval*, Edited by B. Croft and R. Casey, 179-190.

Strzalkowski, T. and Vauthey B., 1992, Information retrieval using robust natural language processing, *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, ACL'92, 104-111.

Sussna, Michael, 1993, Word Sense Disambiguation for Free-text Indexing Using a Massive Semantic Network, *Proceedings of the Second International Conference on Information and Knowledge Management*, CIKM'93, 67-74.

Tanaka, Y., and Yoshida, 1990, Preparation of a Concept Dictionary, *Proceedings of the Second International Congress on Terminology and Knowledge Engineering*, October 1990, 487-495.

Tanimoto, T. 1958, *An Elementary Mathematical Theory of Classification And Prediction*, IBM Internal Report

Tudhope, D. And Taylor, C. 1996, A Unified Similarity Coefficient for Navigating Through Multi-Dimensional Information, *Proceedings of ASIS 1996 Annual Conference.*

Tversky, A., 1977, Features of Similarity. *Psychological Review*, Vol. 84, 327-352.

Tversky, A., and Gati, I., 1978, Studies of Similarity, in *Cognition and Categorization*, edited by Rosch, E. and Lloyd, B. B., Lawrence Erlbaum Associates.

United Nations, 1990, *International Standard Industrial Classification of All Economic Activities*, Third Revision, Statistical Office, New York, Publishing Division, United Nations.

United States, 1987, Office of Management and Budget. *Standard Industrial Classification Manual*, Executive Office of the President, Washington D.C.

Vickery, B. and Vickery, A., 1993, Online Search Interface Design, *Journal of Documentation*, Vol. 49, No. 2, June 1993, 103-187.

Voorhees, E.M., 1994, Query Expansion Using Lexical-Semantic Relations, *Proceedings of the 17th Annual International Conference on Research and Development in Information Retrieval*, SIGIR'94, 61-68.

Voss, D. A. and Driscoll, J.R., 1992, Text retrieval using a comprehensive semantic lexicon, *Proceedings of the First International Conference on Information and Knowledge Management*, CIKM'92, 120-129.

Warren, B. 1978, *Semantic Patterns of Noun-Noun Compounds*, Goteburg.

Watanabe, S, 1985, *Pattern Recognition : Human and Mechanical*, John Wiley & Sons.

Weide, J., Desmarais, N., and Harrington, J., 1992, Optical Product Reviews, *CD-ROM Librarian*, Vol. 7, No. 3, 28-43.

Wendlandt, E.B., and Driscoll, J.R., 1991, Incorporating a Semantic Analysis into a Document Retrieval Strategy, , *Proceedings of the 14th Annual International Conference on Research and Development in Information Retrieval*, SIGIR'91, 270-279

Wilks, Y and Stevenson, M. 1996, *The Grammar of Sense: Is word-sense tagging much more than part-of-speech tagging?*, Working Paper, CS-96-05, University of Sheffield, UK, http://xxx.lanl.gov/abs/cmp-lg/9607028.

Wilks, Y and Stevenson, M. 1997, Sense Tagging: Semantic Tagging with a Lexicon, *Proceedings of the SIGLEX Workshop "Tagging Text with Lexical Semantics"*.

Wolverton, M., 1994, *Retrieving Semantically Distant Analogies*, Ph.D. Thesis, Department of Computer Science, Stanford University.

Wong, S.K.M. and Yao Y.Y., 1992, An Information-Theoretic Measure of Term Specificity, *Journal of the American Society for Information Science*, Vol. 43, No. 1

Woods, A., Fletcher, P. and Hughes, A., 1986, *Statistics in Language Studies*, Cambridge University Press.

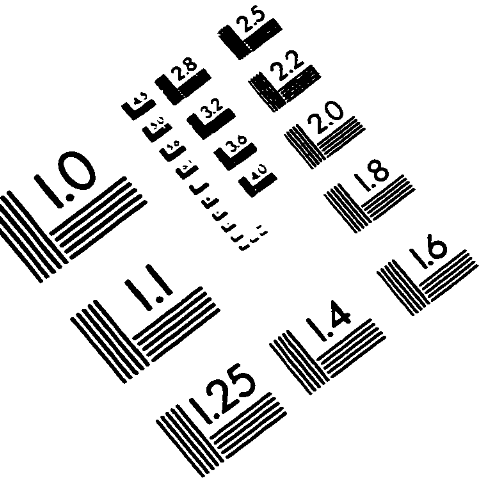
Yarowsky, D. ,1992, Word Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora. In *Proceedings of the 14th International Conference on Computational Linguistics*, Nantes, France, 454-460.

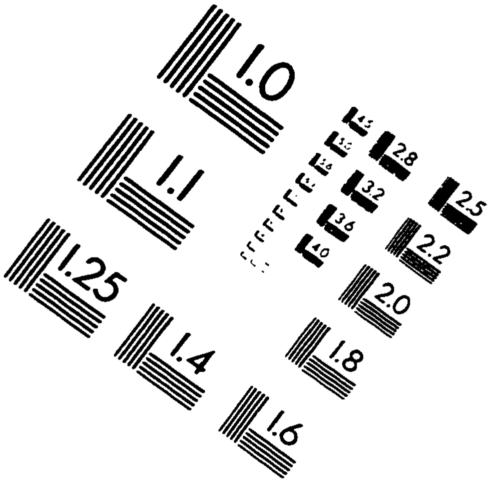
Yarowsky, D., 1993, One Sense per Collocation, *Proceedings of ARPA Workshop on Human Language Technology*, 266-271, March 1993.

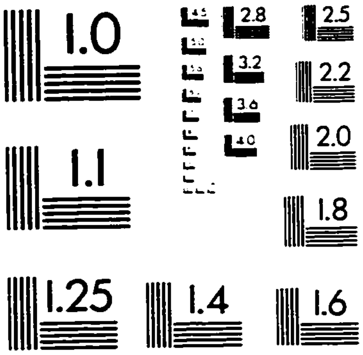
Yarowsky, D., 1995, Unsupervised Word Sense Disambiguation Rivalling Supervised Methods, *In Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, ACL'95, Cambridge, MA, 189-196.

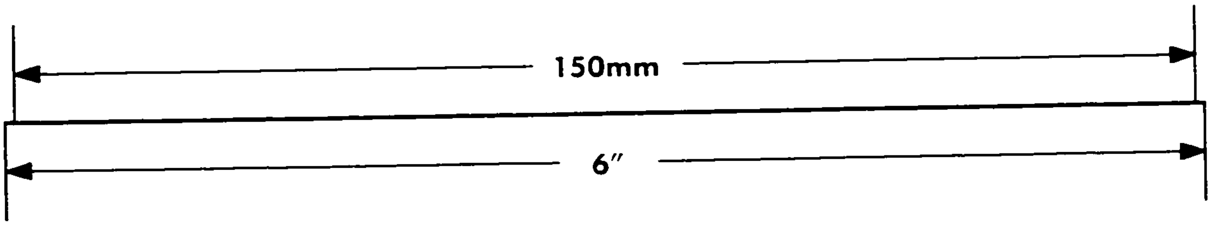
Young, F. and Hamer, R.M., 1987, *Multidimensional Scaling: History, Theory, and Applications*, Lawrence Erlbaum Associates, New Jersey.

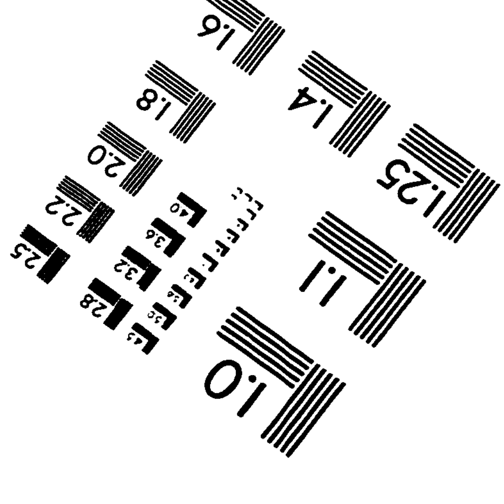
IMAGE EVALUATION TEST TARGET (QA-3)

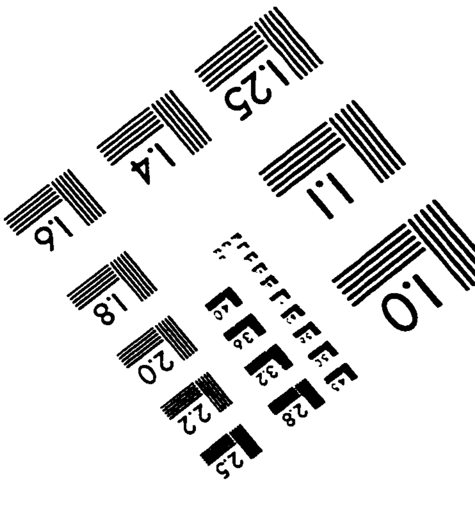












APPLIED IMAGE . Inc
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989