Evaluating the Novelty of Text-Mined Rules Using Lexical Knowledge

Sugato Basu

CS Department
University of Texas
Austin, TX - 78712

sugato@cs.utexas.edu

Raymond J. Mooney

CS Department
University of Texas
Austin, TX - 78712

mooney@cs.utexas.edu

Krupakar V. Pasupuleti

CS Department
University of Texas
Austin, TX - 78712

krupakar@cs.utexas.edu

Joydeep Ghosh

ECE Department
University of Texas
Austin, TX - 78712

ghosh@ece.utexas.edu

ABSTRACT

In this paper, we present a new method of estimating the novelty of rules discovered by data-mining methods using WordNet, a lexical knowledge-base of English words. We assess the novelty of a rule by the average semantic distance in a knowledge hierarchy between the words in the antecedent and the consequent of the rule – the more the average distance, more is the novelty of the rule. The novelty of rules extracted by the DiscoTEX text-mining system on Amazon.com book descriptions were evaluated by both human subjects and by our algorithm. By computing correlation coefficients between pairs of human ratings and between human and automatic ratings, we found that the automatic scoring of rules based on our novelty measure correlates with human judgments about as well as human judgments correlate with one another.

General Terms

Text mining, interesting rules, novelty, semantic distance, knowledge hierarchy, WordNet

1. INTRODUCTION

A data-mining system may discover a large body of rules; however, relatively few of these may convey useful new knowledge to the user. Several metrics for evaluating the "interestingness" of mined rules have been proposed [1, 7]. These metrics can be used to filter out a large percentage of the less interesting rules, thus yielding a more manageable number of higher quality rules to be presented to the user. However, most of these measure simplicity (e.g. rule size), certainty (e.g. *confidence*), or utility (e.g. *support*). Another important aspect of interestingness is *novelty*: does the rule represent an association that is currently unknown. For example, a text-mining system we developed that discovers rules from computer-science job announcements posted to a local newsgroup [18] induced the rule: "SQL $\rightarrow$ database".

A knowledgeable computer scientist may find this rule uninteresting because it conveys a known association. Evaluating the novelty of a rule requires comparing it to an existing body of knowledge the user is assumed to already possess.

For text mining [8, 3, 17], in which rules consist of words in natural language, a relevant body of common knowledge is basic lexical semantics, i.e. the meanings of words and the semantic relationships between them. A number of lexical knowledge bases are now available. WordNet [5] is a semantic network of about 130,000 English words linked to about 100,000 lexical senses (*synsets*) that are interconnected by relations such as antonym, generalization (*hypernym*), and part-of (*holonym*). We present and evaluate a method for measuring the novelty of text-mined rules using such lexical knowledge.

We define a measure of the semantic distance, $d(w_i, w_j)$, between two words based on the length of the shortest path connecting w_i and w_j in WordNet. The novelty of a rule is then defined as the average value of $d(w_i, w_j)$ across all pairs of words (w_i, w_j) where w_i is in the antecedent and w_j is in the consequent of the rule. Intuitively, the semantic dissimilarity of the terms in a rule's antecedent and in its consequent is an indication of the rule's novelty. For example, "beer $\rightarrow$ diapers" would be considered more novel than "beer $\rightarrow$ pretzels" since beer and pretzels are both food products and therefore closer in WordNet.

We present an experimental evaluation of this novelty metric by applying it to rules mined from book descriptions extracted from Amazon.com. Since novelty is fundamentally subjective, we compared the metric to human judgments. We have developed a web-based tool that allows human subjects to enter estimates of the novelty of rules. We asked multiple human subjects to score random selections of mined rules and compared the results to those obtained by applying our metric to the same rules. We found that the average correlation between the scoring of our algorithm and that of the human users, using both raw score correlation (Pearson's metric) and rank correlation (Spearman's metric), was comparable to the average score correlation between the human users. This suggests that the algorithm has a rule scoring judgment similar to that of human users.

2. BACKGROUND

2.1 Text Mining

Traditional data mining algorithms are generally applied to structured databases, but text mining algorithms try to

```
<title>      daring, love
<synopses>   woman
<subject>    romance, historical, fiction
 ->
<comments>   story, read, wonderful
```

Figure 1: DiscoTEX rule mined from Amazon.com "romance" book descriptions

discover knowledge from unstructured or semi-structured textual data, e.g. web-pages. Text mining is a relatively new research area at the intersection of natural language processing, machine learning and information retrieval. Various new useful techniques are being developed by researchers for discovering knowledge from large text corpora, by appropriately integrating methods from these different disciplines. DiscoTEX [18] is one such system, that discovers prediction rules from natural language corpora using a combination of information extraction and data mining. It learns an information extraction system to transform text into more structured data, and this structured data is then mined for interesting relationships.

For our experiments, we have used rules mined by DiscoTEX from book descriptions extracted from Amazon.com, in the "science", "romance" and "literature" categories. DiscoTEX first extracts a structured template from the Amazon.com book description web-pages. It constructs a template for each book description, with pre-defined slots (e.g. title, author, subject, etc.) that are filled with words extracted from the text. DiscoTEX then uses a rule mining technique to extract prediction rules from this template database [19]. An example extracted rule is shown in Figure 1, where the <comments> slot is predicted from the other slots. For our purpose, we only use the filler words in the slot, ignoring the slotnames — in our algorithm, the rule in Figure 1 would be used in the form "daring love woman romance historical fiction → story read wonderful".

2.2 WordNet

WordNet [5] is an online lexical knowledge-base of 130,000 English words, developed at Princeton University. In WordNet, English nouns, adjectives, verbs and adverbs are organized into synonym sets or *synsets*, each representing an underlying lexical concept. A synset contains words of similar meaning pertaining to a common semantic concept. But since a word can have different meanings in different contexts, a word can be present in multiple synsets. A synset contains associated pointers representing its relation to other synsets. WordNet supports many pointer types. The ones we used in our algorithm are: Synonym (*life → existence*), Antonym (*front→ back*), Attribute (*benevolence → good*), Pertainym i.e. morphological relation (*alphabetical → alphabet*), Similar (*unquestioning → absolute*), Cause (*kill → die*), Entailment (*breathe → inhale*), Holonym i.e. part-of (*chapter → text*), Meronym i.e. whole-of (*computer → cpu*), Hyponym i.e. specification (*plant → fungus*), and Hypernym i.e. generalization (*apple → fruit*).

2.3 Semantic Similarity of Words

Several measures of semantic similarity based on distance between words in WordNet have been used by different researchers. Leacock and Chodorow [13] have used the neg-

ative logarithm of the normalized shortest path length as a measure of similarity between two words, where the path length is measured as the number of nodes in the path between the two words and the normalizing factor is the maximum depth in the taxonomy. In this metric, the greater the semantic distance between two words in the WordNet hierarchy, the less is their semantic similarity. Lee et al. [14] and Rada et al. [21] have used conceptual distance, based on an edge counting metric, to measure similarity of a query to documents. Resnick [22] observed that two words deep in the WordNet are more closely related than two words higher up in the hierarchy, both pairs having the same path length (number of nodes) between them. Sussna [26] took this into account in his semantic distance measure that uses depth-relative scaling. Hirst et al. [9] classified the relations of WordNet into the three broad directional categories and used a distance measure where they took into account not only the path length but also the number of direction changes in the semantic relations along the path. Resnick [23] has used an information-based measure instead of path length to measure the similarity, where the similarity of two words is estimated from the information content of the least probable class to which both words belong.

3. SCORING THE NOVELTY OF RULES

3.1 Semantic Distance Measure

We have defined the semantic distance between two words w_i and w_j as:

$$d(w_i, w_j) = Dist(P(w_i, w_j)) + K \times Dir(P(w_i, w_j))$$

where $P(w_i, w_j)$ is the weighted shortest path between w_i and w_j, $Dist(p)$ is the distance along path p according to our weighting scheme, $Dir(p)$ is the number of direction changes of relations along path p, and K is a suitably chosen constant.

The second component of the formula is derived from the definition of Hirst et al. [9], where the relations of WordNet are divided into three direction classes — "up", "down" and "horizontal", depending on how the two words in the relation are lexically related. Table 1 summarizes the direction information for the relation types we use. The more direction changes in the path from one word to another, the greater the semantic distance between the words, since changes of direction along the path reflect large changes in semantic context.

The path distance component of the above formula is based on the semantic distance definition of Sussna [26]. It is defined as the shortest weighted path between w_i and w_j, where every edge in the path is weighted according to the WordNet relation corresponding to that edge, and is normalized by the depth in the WordNet hierarchy where the edge occurs. We have used 15 different WordNet relations in our framework, and we have assigned different weights to different link types, e.g. hypernym represents a larger semantic change than synonym, so hypernym has a higher weight than synonym. The weights chosen for the different relations are given in Table 1.

One point to note here is that Sussna's definition of semantic distance calculated the weight of an edge between two nouns w_i and w_j as the average of the two relations $w_i \to_r w_j$ and $w_j \to_{r'} w_i$ corresponding to the edge, relation r' being the inverse of relation r. This made the

Relation	Direction	Weight
Synonym, Attribute, Similar, Pertainym	Horizontal	0.5
Antonym	Horizontal	2.5
Hypernym, (Member|Part| Substance) Meronym	Up	1.5
Hyponym, (Member|Part| Substance) Holonym, Cause, Entailment	Down	1.5

Table 1: Direction and weight information for the 15 WordNet relations used

```
For each rule in a rule file
    Let A = set of antecedent words,
        C = set of consequent words
    For each word $w_i \in A$ and $w_j \in C$
        If $w_i$ and $w_j$ are not a valid words in WordNet
            Score $(w_i, w_j) \leftarrow$ PathViaRoot$(d_{avg}, d_{avg})$
        Elseif $w_j$ is not a valid word in WordNet
            Score $(w_i, w_j) \leftarrow$ PathViaRoot$(w_i, d_{avg})$
        Elseif $w_i$ is not a valid word in WordNet
            Score $(w_i, w_j) \leftarrow$ PathViaRoot$(d_{avg}, w_j)$
        Elseif path not found between $w_i$ and $w_j$ (in
            user-specified time-limit)
            Score $(w_i, w_j) \leftarrow$ PathViaRoot$(w_i, w_j)$
        Else
            Score $(w_i, w_j) \leftarrow d(w_i, w_j)$
    Score of rule = Average of all $(w_i, w_j)$ scores
Sort scored rules in descending order
```

Figure 2: Rule Scoring Algorithm

semantic distance between two words a symmetric measure. He had considered the noun hierarchy, where every relation between nouns has an inverse relation. But in our framework, where we have considered all the four types of words in WordNet (nouns, adverbs, adjectives and verbs) and 15 different relation types between these words, all of these relations do not have inverses, e.g. the entailment relation has no direct inverse. So, we have used only the weight of the relation $w_i \rightarrow_r w_j$ as a measure of the weight of the edge between w_i and w_j. This gives a directionality to our semantic measure, which is also conceptually compatible with the fact that w_i is a word in the antecedent of the rule and w_j is a word in the consequent of the rule.

3.2 Rule Scoring Algorithm

The scoring algorithm of rules according to novelty is outlined in Figure 2. The algorithm calculates the semantic distance $d(w_i, w_j)$ between two words w_i and w_j, where w_i is in the antecedent and w_j is in the consequent of the rule, based on the length of the shortest path connecting w_i and w_j in WordNet. The novelty of a rule is then calculated as the average value of $d(w_i, w_j)$ across all pairs of words (w_i, w_j) .

The noun hierarchy of the WordNet is disconnected — there are 11 separate hierarchies with distinct root nodes.

The verb hierarchy is also disconnected, with 15 distinct root nodes. For our purpose, following the method of Leacock and Chodorow [13], we have connected the 11 root nodes of the noun hierarchy to a single root node R_{noun} so that a path can always be found between two nouns. Similarly, we have connected the verb root nodes by a single root node R_{verb}. R_{noun} and R_{verb} are further connected to a top-level root node, R_{top}. This connects all the verbs and nouns in the WordNet database. Adjectives and adverbs are not hierarchically arranged in WordNet, but they are related to their corresponding nouns. In this composite connected hierarchy derived from the WordNet hierarchy, we find the shortest weighted path between two words by performing a branch and bound search.

In this composite word hierarchy, any two words are connected by a path. However, we have used 15 different WordNet relations while searching for the path between two words — this creates a combinatorial explosion while performing the branch and bound search on the composite hierarchy. So, for efficient implementation, we have a user-specified time-limit (set to 3 seconds in our experiments) within which we try to find the shortest path between the words w_i and w_j. If the shortest path cannot be found within the time-limit, the algorithm finds a default path between w_i and w_j by going up the hierarchy from both w_i and w_j, using hypernym links, till a common root node is reached.

The function *PathViaRoot* in Figure 2 computes the distance of the default path. For nouns and verbs, the PathViaRoot function calculates the distance of the path between the two words as the sum of the path distances of each word to its root. If the R_{noun} or the R_{verb} node are a part of this path, it adds a penalty term $POSRootPenalty = 3.0$ to the path distance. If the R_{top} node is a part of this path, it adds a larger penalty $TopRootPenalty = 4.0$ to the path distance. These penalty terms reflect the large semantic jumps in paths which go through the root nodes R_{noun}, R_{verb} and R_{top}.

If one of the words is an adjective or an adverb, and the shortest path method does not terminate within the specified time-limit, then the algorithm finds the path from the adjective or adverb to the nearest noun, through relations like "pertainym", "attribute", etc. It then finds the default path up the noun hierarchy, and the PathViaRoot function incorporates the distance from the adjective/adverb to the noun into the path distance measurement.

Some of the words extracted from the rules are not valid words in WordNet e.g. abbreviations, names like Philip, domain specific terms like booknews, etc. We assigned such words the average depth of a word (d_{avg} in Figure 2) in the WordNet hierarchy, which was estimated by sampling techniques to be about 6, and then estimated its path distance to the root of the combined hierarchy by using the PathViaRoot function.

4. EXPERIMENTAL RESULTS

Since novelty is a subjective concept, we performed experiments to compare the novelty judgment of human users to the automatic ratings produced by our algorithm. The objective here is that if the automatic ratings correlate with human judgments about as well as human judgments correlate with each other, then the novelty metric can be considered successful.

```
High score (>8):

romance love heart -> midnight
space astrophysics science astronautics -> apollo
science computers internet -> artificial
front classics -> literature war world
nonlinear -> science physics

Medium score (4-6):

author romance -> characters love
applied science mathematics -> exercises
science physics -> theory
analysis -> science nature
love read -> romance fiction

Low score(<2):

astronomy science -> space
geography -> nature world
mechanics physics -> science
fiction classics -> literature
sea geography -> ocean
```

Figure 3: Examples of rules scored by our novelty measure

	Human - Human Correlation		Algorithm - Human Correlation	
	Raw	Rank	Raw	Rank
Group1	0.284	0.269	0.158	0.113
Group2	0.299	0.282	0.357	0.330
Group3	0.217	0.223	0.303	0.297

Table 2: Results with all subjects

	Human - Human Correlation		Algorithm - Human Correlation	
	Raw	Rank	Raw	Rank
Group1	0.350	0.338	0.187	0.137
Group2	0.412	0.393	0.386	0.363
Group3	0.337	0.339	0.339	0.338

Table 3: Results after removing outliers

4.1 Methodology

For the purpose of our experiments, we took rules that were generated by DiscoTEX from 9,000 Amazon book descriptions, obtained by spidering the `www.amazon.com` website – 2,000 in the "literature" category, 3,000 in the "science" category and 4,000 in the "romance" category. From the total set of rules, we selected a subset of rules that had less than a total of 10 words in the antecedent and consequent of the rule — this was done so that the rules were not too large for human users to rank. Further pruning was performed to remove duplicate words from the rules. For the Amazon.com book description domain, we also created a stoplist of commonly occurring words, e.g. book, table, index, content, etc., and removed them from the rules. There were 1,258 rules in the final pruned rule-set.

We sampled this pruned rule-set to create 4 sets of random rules, each containing 25 rules. 48 subjects were selected for this experiment, most of them being students from Computer Science or Computer Engineering background. The subjects were randomly divided into 4 groups of 12 each. Each group was asked to score one of the rule-sets, using novelty as the scoring criterion. We created a web-interface for the subjects to rank these rules with scores in the range from 0.0 (least novel) to 10.0 (most novel), according to their judgment.

One of the rule-sets was used as a training set, to tune the parameters of the algorithm. The 3 other rule-sets were used as test sets for our experiment. For each of the rule-sets, two types of average correlation were calculated. The first average correlation was measured between the human subjects, to find the correlation in the judgment of novelty between human users. The second average correlation measure was measured between the algorithm and the users in each group, to find the correlation between the novelty scor-

ing of the algorithm and that of the human subjects. We used both Pearson's raw score correlation metric and Spearman's rank correlation metric to compute the correlation measures. We also ran correlation test after removing obvious outliers (i.e. subjects who were negatively correlated to the majority of the other subjects) from the data.

4.2 Results and Discussion

Some of the rules scored by our algorithm are shown in Figure 3. Table 2 shows results using the rankings of all the subjects. Table 3 shows results after removing one outlier from Group1, one from Group2 and two from Group3.

The correlation between the human subjects and the algorithm was low for the first rule-set. For the second and the third rule sets, the algorithm-human correlations are comparable to the human-human correlations. From the results, considering both the raw and the rank correlation measures, we see that the correlation between the human subjects and the algorithm is on the average comparable to that between the human subjects. From Tables 2 and 3, we can see that removing the obvious outliers improves the correlation values. However, the correlation values among the human subjects and between the human subjects and the algorithm are both not very high, even after outlier removal. This is because for some rules, the human subjects differed a lot in their novelty assessment. This is also due to the fact that these are initial experiments, and we are working on improving the methodology. In later experiments, we intend to apply our method to more specific domains, where we can expect human users to agree more in their novelty judgment of rules.

However, it is important to note that it is very unlikely that these correlations are due to random chance — except for the algorithm-human correlation values for Group1, the correlation values considering all subjects are above the minimum significant r at the $p < 0.1$ level of significance, while the correlation values after removing the outliers are all above the minimum significant r at the $p < 0.05$ level of significance, by the t-test.

On closer analysis of the results of Group1, we noticed that this set contained many rules involving proper nouns. Our algorithm currently uses only semantic information from WordNet, so it's scoring on these rules differed from that of human subjects. For example, one rule many users scored

as uninteresting was "ieee society → science mathematics", but since WordNet does not have an entry for "ieee", our algorithm gave the overall rule a high score. Another rule to which some users gave a low score was "physics science nature → john wiley publisher sons", presumably based on their background knowledge about publishing houses. In this case, our algorithm found the name John in the Word-Net hierarchy (synset lemma: disciple of Jesus), but there was no short path between John and the words in the antecedent of the rule. As a result, the algorithm gave this rule a high score. A point to note here is that some names like Jesus, John, James, etc. have entries in WordNet, but others like Sandra, Robert, etc. do not — this makes it difficult to use any kind of consistent handling of names using filters like name lists. In the future, we plan to resolve this problem by using multiple hierarchies with weighting.

In the training rule-set, we had also noticed that the rule "sea → oceanography" had been given a large score by our algorithm, while most subjects in that group had rated that rule as uninteresting. This happened because there is no short path between sea and oceanography in WordNet — these two words are related thematically, and WordNet does not have thematic connections, an issue which is discussed in detail in Section 6.

5. RELATED WORK

Much effort has gone into reducing mined rule-sets by applying both objective and subjective criteria of interestingness. Klemettinen et al. [11] proposed the use of rule templates to describe the structure of relevant rules and constrain the search space. Another notable attempt in using objective measures was by Bayardo and Agrawal [1], who defined a partial order, in terms of both support and confidence, to identify a smaller set of rules that were more interesting than the rest. Sahar [24] proposed an iterative elimination of uninteresting rules, limiting user interaction to a few simple classification questions. Hussain et al. [10] developed a method for identifying exception rules, with the interestingness of a rule being estimated relative to common sense rules and reference rules. In a series of papers, Tuzhilin and his co-researchers [25, 20] argued the need for subjective measures for the interestingness of rules. Rules that were not only actionable but also unexpected in that they conflicted with the existing system of beliefs of the user, were preferred. Liu et al. [15] have further built on this theme, implementing it as an interactive, post-processing routine. They have also analyzed classification rules, such as those extracted from C4.5, defining a measure of rule interestingness in terms of the *syntactic* distance between a rule and a belief. A rule and a belief are "different" if either the consequents of the rule and the belief are "similar" but the antecedents are far apart, or vice versa.

In contrast, in this paper we have analyzed information extracted from unstructured or semi-structured data such as web-pages, and extracted rules depicting important relations and regularities in such data. The nature of rules as well as of prior domain knowledge is quite different from those extracted, say, from market baskets. We have proposed an innovative use of WordNet to estimate the *semantic* distance between the antecedents and consequents of a rule, which is used as an indication of the novelty of the rule. Domain-specific concept hierarchies have previously been used to filter redundant mined rules [6, 4]; however, to

our knowledge they have not been used to evaluate novelty quantitatively, or applied to rules extracted from text data.

6. FUTURE WORK

An important issue for future research is learning the parameters of the algorithm, e.g. the weights of the WordNet relations, and values of K, $POSRootPenalty$ and $TopRoot-Penalty$. These constants are now chosen experimentally. We would like to learn these parameters automatically from training data, by using machine learning techniques. The novelty score could then be adaptively learnt for a particular user and tailored to suit the user's expectations.

Unfortunately, WordNet fails to capture all semantic relationships between words, such as general thematic connections like that between "pencil" and "paper". However, other approaches to lexical semantic similarity, such as statistical methods based on word co-occurrence, can capture such relationships. In these methods, a word is typically represented by a vector in which each component is the number of times the word co-occurs with another specified word within a particular corpus. Co-occurrence can be based on appearing within a fixed-size window of words, or in the same sentence, paragraph, or document. The similarity of two words is then determined by a vector-space metric such as the cosine of the angle between their corresponding vectors [16]. In techniques such as *Latent Semantic Analysis* (LSA) [2], the dimensionality of word vectors is first reduced using *singular value decomposition* (SVD) in order to produce lexical representations with a small number of highly-relevant dimensions. Such methods have been shown to accurately model human lexical-similarity judgments [12]. By utilizing a co-occurrence-based metric for $d(w_i, w_j)$, rules could be ranked by novelty using statistical lexical knowledge.

In the end, some mathematical combination of WordNet and co-occurrence based metrics may be the best approach to measuring lexical semantic distance. To the extent that the names of relations, attributes, and values in a traditional database are natural-language words (or can be segmented into words), our approach could also be applied to traditional data mining as well as text mining. Finally, the overall interestingness of a rule might be best computed as a suitable mathematical combination of novelty and more traditional metrics such as confidence and support. Some of the problems we faced while using a "generic" ontology like WordNet can be solved by using multiple domain-specific concept hierarchies, if available.

7. CONCLUSION

The main contribution of this paper is that we have introduced a new approach for measuring the novelty of rules mined from text data, based on the lexical knowledge in WordNet. We have also introduced a novel method of quantitatively assessing interestingness measures for rules, based on average correlation statistics, and have successfully shown that the automatic scoring of rules based on our novelty measure correlates with human judgments about as well as human judgments correlate with each other.

8. ACKNOWLEDGMENTS

9. REFERENCES

[1] R. J. Bayardo Jr. and R. Agrawal. Mining the most interesting rules. In *Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining (KDD-99)*, pages 145–154, San Diego, CA, August 1999.

[2] S. C. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41:391–407, 1990.

[3] R. Feldman, editor. *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI-99) Workshop on Text Mining: Foundations, Techniques and Applications*, Stockholm, Sweden, August 1999.

[4] R. Feldman and I. Dagan. Knowledge discovery in textual databases (KDT). In *Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95)*, pages 112–117, 1995.

[5] C. D. Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, 1998.

[6] J. Han and Y. Fu. Discovery of multiple-level association rules from large databases. In *Proceedings of the 21st International Conference on Very Large Databases (VLDB-95)*, pages 420–431, Zurich, Switzerland, 1995.

[7] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kauffmann Publishers, San Francisco, 2001.

[8] M. Hearst. Untangling text data mining. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99)*, pages 3–10, College Park, MD, June 1999.

[9] G. Hirst and D. St-Onge. Lexical chains as representations of context for the detection and correction of malapropims. In C. Fellbaum, editor, *WordNet: An Electronic Lexical Database*, chapter 13, pages 305–332. MIT Press, 1998.

[10] F. Hussain, H. Liu, E. Suzuki, and H. Lu. Exception rule mining with a relative interestingness measure. In *Proceedings of Pacific Asia Conference on Knowledge Discovery in DataBases (PAKDD-2000)*, pages 86–97, April 2000.

[11] M. Klemettinen, H. Mannila, P. Ronkainen, H. Toivonen, and A. Verkamo. Finding interesting rules from large sets of discovered association rules. In *Proceedings of The Third International Conference on Information and Knowledge Management (CIKM-94)*, pages 401–407, 1994.

[12] T. K. Landauer and S. T. Dumais. A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104:211–240, 1997.

[13] C. Leacock and M. Chodorow. Combining local context and WordNet similarity for word sense identification. In C. Fellbaum, editor, *WordNet: An Electronic Lexical Database*, chapter 11, pages 265–284. MIT Press, 1998.

[14] J. H. Lee, M. H. Kim, and Y. J. Lee. Information retrieval based on a conceptual distance in IS-A heirarchy. *Journal of Documentation*, 49(2):188–207, June 1993.

[15] B. Liu, W. Hsu, L.-F. Mun, and H. Lee. Finding interesting patterns using user expectations. *IEEE Transactions on Knowledge and Data Engineering*, 11(6):817–832, 1999.

[16] C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, 1999.

[17] D. Mladenić, editor. *Proceedings of the Sixth International Conference on Knowledge Discovery and Data Mining (KDD-2000) Workshop on Text Mining*, Boston, MA, August 2000.

[18] U. Y. Nahm and R. J. Mooney. A mutually beneficial integration of data mining and information extraction. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence (AAAI-2000)*, pages 627–632, Austin, TX, July 2000.

[19] U. Y. Nahm and R. J. Mooney. Mining soft-matching rules from textual data. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI-2001)*, Seattle, WA, 2001.

[20] B. Padmanabhan and A. Tuzhilin. A belief-driven method for discovering unexpected patterns. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98)*, pages 94–100, 1998.

[21] R. Rada, H. Mili, E. Bicknell, and M. Blettner. Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man and Cybernetics*, 19(1):17–30, 1989.

[22] P. Resnick. WordNet and distribution analysis: A class-based approach to lexical discovery. In *Statistically-Based Natural-Language-Processing Techniques: Papers from the 1992 AAAI Workshop*. AAAI Press, 1992.

[23] P. Resnick. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI-95)*, pages 448–453, 1995.

[24] S. Sahar. Interestingness via what is not interesting. In *The Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 332–336, August 1999.

[25] A. Silberschatz and A. Tuzhilin. What makes patterns interesting in knowledge discovery systems. *IEEE Transactions on Knowledge and Data Engineering*, 8(6):970–974, 1996.

[26] M. Sussna. Word sense disambiguation for free-text indexing using a massive semantic network. In *Proceedings of The Second International Conference on Information and Knowledge Management (CIKM-93)*, pages 67–74, 1993.