IDEA GROUP PUBLISHING

701 E. Chocolate Avenue, Suite 200, Hershey PA 17033, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com

ITP5281

This paper appears in *Managing Modern Organizations Through Information Technology*, Proceedings of the 2005 Information Resources Management Association International Conference, edited by Mehdi Khosrow-Pour. Copyright 2005, Idea Group Inc.

Web Mining for Business Intelligence: Discovering Novel Association Rules from Competitors' Websites

Xin Chen and Yi-fang Brook Wu

Infor. Systems Dept, New Jersey Institute of Technology, Newark, NJ 07102, USA, {xin.chen, wu@njit.edu}

ABSTRACT

The Web has offered companies a convenient way to publish their information and conduct business with customers and partners, and it also opens an opportunity to companies to acquire knowledge of their competitors. Such knowledge is critical for a company to define its business strategies and to establish a network with business partners. This paper proposes a content web mining technique that discovers novel association rules among noun phrases extracted from web pages on one's competitors' websites. Novelty of an association rule is measured as the distance between the antecedent and the consequent of the rule in the background knowledge, which is developed from documents on one's own website are. Incorporating background knowledge into the web mining process enables us to discover previously unknown but potential useful patterns. A running example demonstrates that the novelty prediction accuracy is high in terms of correlating with human judgments.

INTRODUCTION

Companies have been publishing various types of information online. However, finding valuable information from competitors' websites is not an easy task because (1) the number of web pages is too large for human to seek such information manually, and (2) patterns and relationships between entities cannot be found without analyzing web pages collectively. Information searching tools (e.g. search engines) can be designed to overcome the first difficulty. However, in the circumstances of finding unexpected information, users' information needs are unknown in advance. Search engines also lack of the ability to analyze the retrieved documents to discover hidden patterns. The ability to gain business intelligence by exploiting searching tools is very limited.

Web mining techniques aim at discovering knowledge from the content of web pages, system logs and the link structures among web pages. Web content mining is a special case of text mining. When competing with its competitors and cooperating with its partners, a company needs to know what it does not know. Under such circumstances, the discovered patterns tend to be more useful if they are previously unknown or unexpected. However, most text mining techniques follow a document-oriented view, in which the discovered knowledge is solely determined by the target document collection, while users' expectation or background is seldom considered. Another problem with text mining techniques is that the discovered patterns can easily reach tens of thousands, and it is painful for the user to sift useful ones.

There is a need to incorporate users' background knowledge into the text mining process. User-oriented text mining techniques can not only filter out uninteresting patterns, but also identify novel (and potentially useful) patterns for a particular user. For the problem of discovering unknown knowledge of competitors, we can reasonably assume that a company's own website contains information that is known to the company. This paper presents a novel approach to discovering previously unknown knowledge from competitors' websites with the consideration of the background knowledge extracted from one's own website. A running example is given to show the effectiveness of the proposed method.

BACKGROUND KNOWLEDGE DEVELOPMENT

Liu et al (2001) introduced an approach to discovering unexpected information from competitors' websites, but the information is limited to terms, web pages and hyperlinks, rather than association rules explored in this study. We borrow the same idea of using the company's own website to develop the background knowledge, and employing such knowledge to discover novel association rules. A key word space, including a concept hierarchy developed inside, is built to represent the background knowledge.

Key Words Extraction

Key words are content bearing and non-functional words. A word is selected as a key word if it does not appear in a stop-word (commonly used words, such as *a, the, his*, etc.) list. All key words are converted to their base forms. It is necessary to address the distinction between *key words* used in this paper and *keywords* used in academic papers. In this paper, a *key word* refers to a single word that appears in a document but not in the stop-word list. In academic articles, *keywords* are a few phrases assigned to identify the main topics of an article or the major categories an article belongs to. From now on, we will explicitly use *phrase* to refer to a term that contains one or more words.

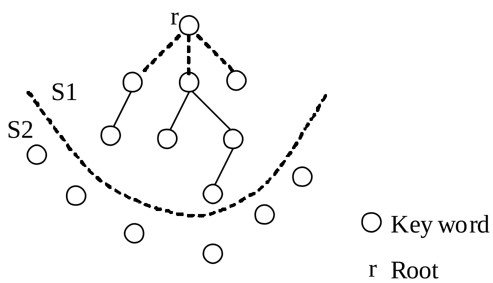
Concept Hierarchy Development

In Information Retrieval, the generality and specificity of terms are measured by their document frequency (DF). The more documents a term occurs in, the more general it is. Forsyth and Rada (1986) introduce the use of DF to derive a multi-level structure that has general terms on top of specific terms. Sanderson and Croft (1999) apply this idea to build and present concept hierarchies derived from text by using subsumption to create a topic hierarchy. Wu (2001) developed a revised subsumption called probability of co-occurrence analysis (POCA). It is defined as $P(X|Y) > P(Y|X)$, $P(X|Y) >= N$, where $0 < N <= 1$. If a term pair (X, Y) fulfills the above set of inequalities, X is the parent of Y. A document frequency threshold (*df*) is also defined to remove unimportant or uncommon key words that appear in less than *df* documents. The POCA technique is used to develop the concept hierarchy from key words extracted from background documents (the background knowledge is shown in figure 1). S1 contains key words that are included in the hierarchy, and S2 contains key words that do not satisfy the co-occurrence probability or the *df* constraints. A virtual key word, *r*, is introduced as the root to connect all first-level key words in the hierarchy.

PRE-PROCESSING OF TARGET DOCUMENTS

Target documents are collected from competitors' websites. Preprocessing of target documents includes extracting document features and selecting important document features.

Figure 1. Background Knowledge



Feature Extraction

Noun phrases are extracted because they are more descriptive than single words, and using phrases as features can also avoid mining such rules resulted from phrase usage (e.g. *Wall -> Street*). A simplified rule-based noun phrase extractor (Brill, 1992) is implemented. To assign the initial Part-Of-Speech (POS) tag, we use a simplified WordNet database (Fellbaum, 1998), which consists of words divided into four categories (*noun, verb, adjective*, and *adverb*) and the number of senses of each word in one of the categories. The initial POS tag for a word is determined by selecting the category with the maximum number of senses. If a word is found in more than one category, it is marked as a multi-tag word. The parser disambiguates a multi-tag word by examining its previous *n* (2~4) tokens against a list of manually defined syntactic rules. For example, "*hit*" can be either a noun or a verb. If the previous word is a determiner (*the, a, this*, etc), it will be tagged as a noun. Additional heuristics (e.g. the ending of a word) are also used to determine the right tag. Noun phrases are identified by selecting the POS sequences that are of interests. The current sequence pattern is defined as [A|N] N, where A refers to Adjective and N refers to Noun. The pattern defines a base noun phrase that consists of a head N and none or more modifiers [A|N].

Feature Selection

To reduce the number of features, and more importantly, to select the significant features, we apply TF.IDF (Salton and Buckley, 1988), a common term weighting scheme in information retrieval, for feature selection. The rationale behind TF.IDF weighting scheme is that the more frequently a term appears in a document, the more important the term is in that document; a term becomes less important when it occurs in more documents in the collection. It is formally defined as $w_i = f_i \cdot \log\left(N / df_i \right)$, where w_i is the weight of the i_{th} term in a document, *fi* is the term frequency, *N* is the total number of documents in the collection, and *dfi* is the document frequency of the term. For each document, noun phrases with a weight smaller than a defined threshold are removed.

NOVEL ASSOCIATION RULES MINING

The knowledge to be discovered from competitors' websites is in the form of association rules. In Data Mining research, both objective and subjective measures have been proposed to identify interesting rules (Silberschatz and Tuzhilin, 1995; Piatesky-Shapiro, 1991). Objective measures rely on the attributes (*length, support*, etc.) of rules and data collection, and are not sufficient for identifying interesting patterns. Subjective measures require users to express what is expected and unexpected, which is difficult to most users. Basu et al (2001) use WordNet to evaluate the novelty of association rules by calculating the semantic distance between two words in WordNet. WordNet, however, is a general lexical database and does not differentiate users with different backgrounds. In this study, the key word space with a concept hierarchy inside is used to measure the novelty of discovered association rules.

Association Rules Mining Among Noun Phrases

Association rules mining over basket data was first introduced by (Agrawal et al, 1993). The goal is to generate all significant association rules between items in the transaction database. Because each target document has been converted to a vector of noun phrases, it can be viewed as a transaction, and the noun phrases can be viewed as the sale items. The association rule mining problem is usually decomposed into two subproblems: frequent itemset identification and rule generation from frequent itemsets. Frequent itemsets are combinations of items that have a greater *support* than the specified minimum threshold. For every frequent itemset, its items are partitioned into two parts: one for the antecedent and one for the consequent. *Confidence* is calculated according to the *support* values of the two parts and the entire itemset. We implement the standard APRIORI algorithm to generate association rules satisfying the given *support* and *confidence* constraints.

Novelty Measurement

Novelty of an association rule is measured by the distance between the antecedent and the consequent of the rule in the background knowledge. For example, given a rule [A, B] -> [C, D], its novelty is calculated as *average(d(A,C), d(B,C), d(A,D), d(B,D))*, where *d(X,Y)* is the distance between item *X* and *Y*. The semantic distance between two key words in the background knowledge is measured from two perspectives: occurrence similarity and connection similarity. The former measures how similar or interchangeable two key words are. The more often they co-occur, the more similar or interchangeable they are. Connection similarity measures how they are connected to other common key words. It captures the relationship between two key words even if they do not co-occur but both are connected to other common key words. The length of the path between the two keywords, namely hierarchy distance, can be used to measure the connection similarity. Figure 2 shows the occurrence similarity and the connection similarity of *X* and *Y*.

We define the semantic distance between two key words *X* and *Y* in the background knowledge as follows:

$$D(X,Y) = \left(1 - \frac{P(XY)}{P(X \cup Y)}\right) \cdot \frac{d(X,Y)}{2(H+2)},$$

where *D(X, Y)* is the semantic distance between *X* and *Y*, *P(XY)* is the probability that *X* and *Y* co-occur, and *P(XUY)* is the probability that *X*

Figure 2. Occurrence Similarity and Connection Similarity

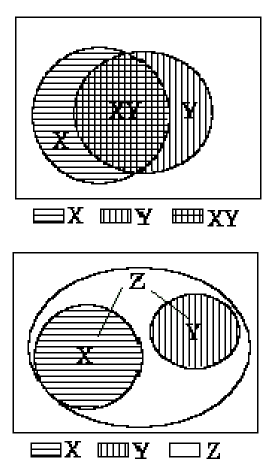
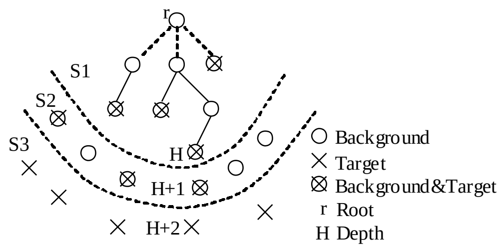


Figure 3. Hierarchy Distance Calculation



or Y occurs. $\frac{P(XY)}{P(X \cup Y)}$ is the occurrence similarity of X and Y. $d(X, Y)$ is the hierarchy distance between X and Y, which equals to length of the shortest path between X and Y in the hierarchy. H is the depth of the hierarchy, and is introduced to normalize the hierarchy distance. $2(H+2)$ is the maximum hierarchy distance between two key words.

First we define two types of key words – background key words and target key words. The former refer to those key words that appear in background documents, and the latter are the key words that occur in target documents. In figure 3, they are shown as circles and crosses respectively. Three areas, S1, S2, and S3, contain different types of key words. S1 and S2 contain all background key words, and S3 contains target key words that cannot be found in background documents. Target key words can also fall into area S1 and S2. H is the depth of the hierarchy, which is the number of words in the longest path from the root r to any leaf in the hierarchy. We now define the hierarchy distance $d(X, Y)$ between two key words X and Y in different areas.

There are two basic conditions: X and Y are both found in the hierarchy, and otherwise. In the first condition, the paths between X and Y in the hierarchy are identified, and the shortest one is selected and its length is assigned as the distance between X and Y. When the selected path includes the root, we add 1 to the length because the root is not an actual key word. In the second condition, X and Y are not both present in the hierarchy, so there is no real connection between them. In such cases, we define $d(X, Y)$ as the sum of $d(X, r)$ and $d(Y, r)$. The hierarchy distance between a key word W and the root r is calculated as (1) the length of the path between W and r if W is in S1, or (2) H+1 if W is in S2, or (3) H+2 if W is in S3. The maximum hierarchy distance is reached when two key words are both in S3. The calculation is summarized in table 1.

We have presented the details of the proposed methodology of discovering novel association rules from documents on competitor's websites. Below we give a running example to show the effectiveness of the proposed algorithm.

A RUNNING EXAMPLE

Experimental Settings

We selected 10 websites of Information Systems departments as competitors by issuing a query "information systems department site:edu" to Google and selecting the top 10 IS related departments. The websphinx crawler (Miller and Bharat, 1998) was used to download web pages from all websites. A total of 512 background documents and 1,422 target documents were collected. 12,945 key words were extracted from the background documents. A minimum document frequency of 6 and a probability threshold of 0.8 (same as the threshold N in Sanderson's and Wu's studies) were chosen for hierarchy development. This setting results in a hierarchy with a depth of 13. Noun phrases were extracted from target documents and their TF.IDF values were calculated for each document. The low 20% noun phrases were removed from each document.

Table 1. Hierarchy Distance Calculation

X \ Y	S1	S2	S3
S1	Len(X-Y) or Len(X-Y)+1	Len(X-r)+(H+1)	Len(X-r)+(H+2)
S2	Len(Y-r)+(H+1)	2(H+1)	(H+2)+(H+1)
S3	Len(Y-r)+(H+2)	(H+2)+(H+1)	2(H+2)
Note: Len(X-Y) is the length of the shortest path between X and Y in the hierarchy			

Table 2. Discovered Rules at Each Novelty Level

Novelty	# of Rules	Example
5	33	[Thomas Jeneklasen] -> [Photography]
4	52	[Trademarks] -> [Karl Eller Center]
3	796	[Admission] -> [Assistantships]
2	271	[Final Exam] -> [Office Hours]
1	3,770	[Phone] -> [Office]

Discovered Rules

The *support* and *confidence* constraints were set to 1% and 60% respectively. A low support was chosen to increase the recall of useful rules. In this running example, we only generated two-item rules (rules with two noun phrases), but it was not difficult to generate rules with more items. Novelty of rules was calculated and normalized to 1 to 5. A total of 4,922 association rules were discovered. The number of rules at each novelty level is shown in table 2. An example is also given for each novelty level.

To evaluate the prediction accuracy of the novelty score, we conducted a user evaluation.

Subjective Evaluation

Eight subjects who have been in our department for at least one year were invited to participate in this evaluation. Because it is not feasible to evaluate all discovered rules, we randomly selected 16 rules from each novelty level (1~5) to create a sample of 80 rules. The subjects were first asked to spend 30 to 50 minutes browsing our department's website and then evaluate the novelty of rules in a 5-point Likert scale (1 for the lest and 5 for the most).

The Kappa Statistic K and the Kendall's Coefficient of Concordance W is tested to ensure the level of intersubject agreement. They are all significant at $p<0.005$ level, and indicate that intersubject agreement is sufficient for further investigation into correlation between human judgments and system predictions. We calculated both Pearson's raw score correlation and the Spearman's rank correlation between the system prediction and the human ratings. Correlations between human subjects for the raw score of novelty and the rank order of rules are 0.315 and 0.270 respectively, and correlations between human subjects and system prediction for the raw score of novelty and the rank order of rules are 0.444 and 0.415 respectively.

Correlation between the system predictions and the human subjective ratings is comparable to, or even slightly better than, that between the human subjects. Because only 80 out of 4,922 rules were selected for evaluation, we conducted a significance test to see if the correlation is also valid for the population. All the t-tests are above the minimum significant r at the $p<0.01$ level of significance. This leads to the conclusion that the correlations are not due to the random chance. The results of the correlation tests indicate that the system accurately predicts the novelty of association rules.

CONCLUSION

In e-business environment, the knowledge about competitors is critical for a company to set up appropriate competing strategies and to establish networks with partners. We illustrated the application of the proposed method to knowledge discovery from competitors' websites.

The experimental result is promising. The system successfully predicted the novelty of extracted association rules in terms of correlating with human ratings at a level comparable to that between human subjects. Using the novelty measure to rank the rules can eliminate about 80% of the uninteresting rules if a mid-point novelty value is selected.

The proposed method emphasizes on discovering unknown knowledge, though companies might also be interested in knowing the similar activities being taken by its competitors. In such cases the low-novelty association rules might be of interests. One limitation of this study is that the evaluation was conducted in educational institutions. We are planning to conduct similar experiments with commercial organizations. In addition to the web pages on a company's websites, documents from other sources, such as company reviews and stock news, could also be used as background documents or target documents.

REFERENCES

Agrawal, R., Imilienski, T. and Swami, A. (1993). Mining Association Rules between Sets of Items in Large Datasets. Proceedings of the ACM SIGMOD International Conference on the Management of Data, pp 207-216, 1993

Basu, S., Mooney, R. J., Pasupuleti, K. V. and Ghosh, J. (2001). Evaluating the Novelty of Text-mined Rules Using Lexical Knowledge. Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2001) (short paper), pp. 233-238, San Francisco, CA, August 2001

Brill, E. (1992). A Simple Rule-based Part of Speech Tagger. Proceedings of ANLP-92, 3rd Conference on Applied Natural Language Processing

Fellbaum, C. D. (1998). Wordnet: An Electronic Lexical Database. MIT Press, Cambridge, MA, 1998

Forsyth, R. and Rada, R. (1986). Adding An Edge in Machine Learning: Applications in Expert Systems and Information Retrieval. (pp. 198-212), Ellis Horwood Ltd

Liu, B., Ma, Y. and Lee, R. (2001). Analyzing the Interestingness of Association Rules from the Temporal Dimension. IEEE International Conference on Data Mining (ICDM-2001), Nov 29 - Dec 2, 2001, Silicon Valley, CA

Miller, R. C. and Bharat, K. (1998). Sphinx: A Framework for Creating Personal Site-specific Web Crawlers. In Proceedings of WWW7, Brisbane Australia, April 1998

Piatesky-Shapiro, G. (1991). Discovery, Analysis, and Presentation of Strong Rules. In G. Piatetsky-Shapiro, and W. J Frawley (eds), Knowledge Discovery in Databases, AAAI/MIT Press, 1991, pg, 231-233

Salton, G. and Buckley, C. (1988). Term Weighting Approaches in Automatic Text Retrieval. Information Processing and Management, vol. 24, no. 5, pp. 513—523, 1988

Sanderson, M. and Croft, B. (1999). Deriving Concept Hierarchies from Text. Proceedings of the 22nd annual international ACM SIGIR Conference on Research and Development in Information Retrieval. 206-213

Silberschatz, A. and Tuzhilin, A. (1995). On Subjective Measures of Interestingness in Knowledge Discovery. In Proceedings of the First International Conference on Knowledge Discovery and, Data Mining, 1995, pg, 275-281

Wu, Y. (2001). Automatic Concept Organization: Organizing Concepts from Text through Probability of Co-occurence Analysis (POCA). PhD thesis, 2001