Personalized Knowledge Discovery: Mining Novel Association Rules from Text *

Xin Chen †, Yi-Fang Wu †

Abstract

This paper presents a methodology for personalized knowledge discovery from text. It derives a user's background knowledge from his/her background documents, and exploits such knowledge to evaluate the novelty of discovered knowledge in the form of association rules by measuring the semantic distance between the antecedent and the consequent of a rule in the background knowledge. The experiment results show that the proposed user-oriented novelty measure is highly correlated with the human subjective rule novelty and usefulness ratings. It outperforms seven major objective interestingness measures and the WordNet novelty measure for identifying novel and useful rules.

Keywords: Text Mining, Personalization, Association Rule Mining, Novelty, Interestingness

1. Introduction

Data mining tools tend to produce a huge number of patterns which makes it difficult for users to find interesting and useful ones quickly and easily. In a study conducted by Stanford University, the association rule mining algorithm generated over 20,000 rules from a subset of the census data containing about 30,000 records. Most of the rules are not useful, and those "that came out at the top, are things that were obvious" [4]. In text mining, the problem becomes even more critical because of the large number of documents available and the high dimensionality of textual data.

Both objective and subjective measures have been proposed to evaluate the interestingness of discovered patterns [8, 10, 11, 13]. However, objective measures alone are insufficient, because they rely only on the characteristics (surface features) of the patterns and the underlying data collection without considering users' knowledge and interests. One can generate a large number of rules that are interesting "objectively" but of little interest to the user [6]. Subjective measures, such as unexpectedness (a pattern is interesting if it is "surprising" to the user) and actionability (a pattern is interesting if the user can act on it to his/her benefit) [13], assess the interestingness of patterns from the users' perspective, but they require explicit expressions of users' subjective opinions (expectation/unexpectation) in order to perform the comparison. In practice it is difficult or even nearly impossible for users to do so, especially before the discovered patterns are presented to them.

This paper presents a text mining technique that discovers novel association rules from documents for a particular user. The system derives a user's background knowledge implicitly from a set of documents that are already known to the user (i.e. background documents). It then applies the background knowledge to retrieve relevant (target) documents from a large corpus and to evaluate the novelty of the association rules discovered from target documents. The experiment results show that the proposed user-oriented novelty measure is highly correlated with the human subjective rule novelty and usefulness ratings. It outperforms seven major objective interestingness measures and the WordNet novelty measure [2] for identifying novel and useful association rules.

2. Deriving User's Background Knowledge

Subjective measures evaluate rule interestingness by comparing them to the explicit expressions of the user's expectation/unexpectation of the result [7] or his/her beliefs [9, 10]. These explicit expressions are difficult to obtain in practice. Also, in text mining tasks, the number of attributes is so large that user specifications, if any, have very limited coverage. A technique that can implicitly capture the user's knowledge or interests is needed.

It is reasonable to assume that a user's knowledge can be reflected by the documents s/he has already read, so the documents known to a user can be a good source to derive his/her background knowledge. The following sub-sections describe the process of constructing a concept hierarchy to model a user's background knowledge from his/her background documents.

2.1 Keywords Extraction. Keywords are content bearing and non-functional words extracted from the background documents. A word is selected as a keyword if it does not appear in a pre-defined stop-word (commonly used words, such as *a*, *the*, *his*, etc.) list. All keywords are converted to their base forms, and are indexed into an inverted list, in which each node consists of a keyword and a list of documents the keyword occurs in.

It is necessary to address the distinction between *keyword* used in this paper and *keywords* used in most academic papers. In this paper, a keyword 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 that the

* Partial support for this research was provided by the United Parcel Service Foundation; the National Science Foundation under grants DUE-0226075, DUE-0434581 and DUE-0434998, and the Institute for Museum and Library Services under grant LG-02-04-0002-04

† Information Systems Department, New Jersey Institute of Technology, Newark, NJ 07102. Email: {xc7, wu}@njit.edu

author(s) assign to an article to identify the main topics of the article or the major categories the article belongs to. From now on, we will explicitly use phrase to refer to a term that consists of one or more words.

2.2 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 introduce the use of DF to derive a multi-level structure that has general terms on top of specific terms [5]. Sanderson and Croft apply this idea to build and present concept hierarchies derived from text by using subsumption to create a topic hierarchy [12]. However, in some cases, subsumption might yield term pairs *(X, Y)* where *X* does not subsume *Y*. To overcome this problem, Wu developed a revised subsumption called probability of co-occurrence analysis (POCA) [15], which states that *X* is the parent of *Y* if $P(X|Y) > P(Y|X)$, $P(X|Y) >= N$, where $0 < N <= 1$. A document frequency threshold (*df*) is also defined to remove keywords that appear in less than *df* documents. The POCA technique is used to develop the concept hierarchy in this study.

2.3 The Keyword Space. After keywords are extracted from background documents and the concept hierarchy is developed, a keyword space that represents the user's background knowledge is constructed (shown in Figure 1). In the keyword space, area S1 contains keywords that are included in the concept hierarchy, and area S2 contains keywords that are not (keywords that do not satisfy the *df* constraint). A virtual keyword *r* is introduced as the root to connect all first-level keywords in the hierarchy.

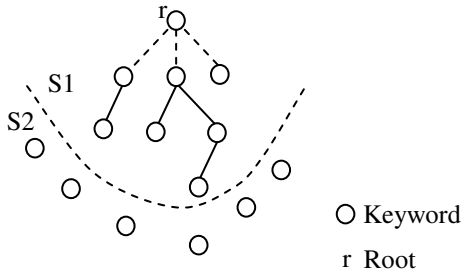

Figure 1: Background Knowledge Keyword Space

3. Knowledge Discovery

The knowledge to be discovered is in the form of association rules, which are mined from target documents retrieved from a large corpus.

3.1 Target Document Retrieval. Target document selection can be viewed as a document retrieval process. Standard information retrieval methods, such as similarity measures and Boolean keyword search, can be used to retrieve relevant documents from a corpus.

3.2 Feature Extraction. Noun phrases are extracted from target documents as document features. Our part-of-speech tagger is a revised version of the widely used Brill tagger [3]. It was trained on two corpora, the Penn Treebank Tagged Wall Street Journal Corpus and the Brown Corpus. After all the words in the document are tagged, noun phrases are extracted by selecting the tokens whose POS sequence matches the pre-defined patterns. The current sequence pattern is defined as A* N+, where A refers to Adjective, N refers to Noun, * means none or more occurrences, and + means one or more occurrences.

The TF.IDF term weighting scheme is applied to select significant noun phrases from each target document. TF is the number of occurrences of a term in a document. DF is the number of documents in which a term occurs, and IDF (inverse document frequency) is a logarithm function of DF. The rationale behind TF.IDF weighting is that the more frequently a term appears in a document, the more important the term is to that document; while a term becomes less important when it occurs in more documents in the collection.

3.3 Association Rule Mining. After feature extraction and selection, target documents are converted into structured vectors of noun phrases. The standard APRIORI algorithm [1] is executed to identify the frequent noun phrase sets and the association rules among noun phrases.

4. Novelty Evaluation

The number of discovered association rules is usually too large for a user to look for interesting rules quickly and easily. Basu et al. use the WordNet database to measure the novelty of association rules by calculating the semantic distance between two words in WordNet [2]. WordNet, however, is a general lexical database and does not differentiate users with different backgrounds. In this study, the keyword space containing a concept hierarchy developed from the user's background documents is used to measure the novelty of association rules. Because the background knowledge is derived from the documents a user has provided, the generated novelty measure is user customized, so the proposed novelty measure is called user-oriented novelty measure.

4.1 The User-Oriented Novelty Measure. The user-oriented novelty of an association rule is defined as the distance between the antecedent and the consequent of the rule in the background knowledge. The distance between two itemsets is defined as the average of the distances between all term pairs, each of which consists of one term from the antecedent and one from the consequent of the rule. 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 semantic distance between item *X* and *Y*. The semantic distance between two keywords in the

background knowledge is measured from two perspectives – occurrence distance and connection distance.

4.2 Occurrence Distance. Occurrence distance measures how distinct the occurrences of two keywords are in the background documents. Given two key words *X* and *Y*, the more often they co-occur, the less the occurrence distance is. A larger distinction in the occurrences of *X* and *Y* indicates a less strength of association between *X* and *Y*. Figure 2 shows the occurrence distance between *X* and *Y*.

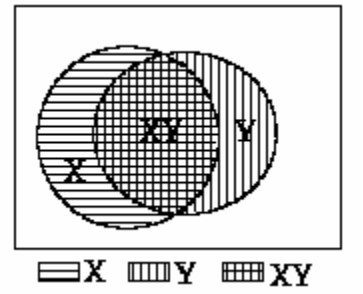

Figure 2: Occurrence Distance

Given the probability that *X* and *Y* co-occur *P(XY)*, the distinction between the occurrences of *X* and *Y* is *P(XUY)-P(XY)*, where *P(XUY)* is the probability of the joint occurrence of *X* and *Y*. If we normalize the occurrence distance by the joint occurrence, the occurrence distance can be denoted as

$$D_o(X, Y) = (P(XUY)-P(XY))/P(XUY)$$

$$= 1-P(XY)/P(XUY).$$

When two keywords do not co-occur, their occurrence distance is 1; when they have the exact same occurrences, their occurrence distance is 0.

4.3 Connection Distance. Connection distance measures the strength of the connection between two keywords in the concept hierarchy. It is possible that two keywords may still have a certain relationship, even if they do not co-occur in the background documents. Figure 3 shows the distance between *X* and *Y* through the connection with the common keyword *Z*.

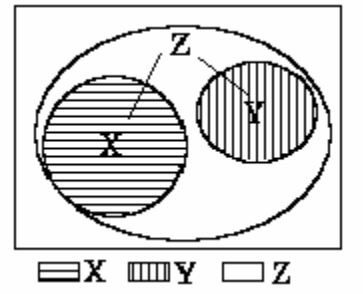

Figure 3: Connection Similarity

The connection distance is defined as the length of the path connecting *X* and *Y* in the hierarchy. The longer the path is, the weaker the connection is. It is normalized by the maximum hierarchy distance between any two keywords.

For easy explanation, we classify keywords into two categories: background keywords and target keywords. The former refers to those keywords that appear in background documents, and the latter refers to the keywords that occur in target documents. The two types of keywords are overlapped, since keywords can appear in both background and target documents. The keywords and their possible locations in the keyword space are shown in Figure 4.

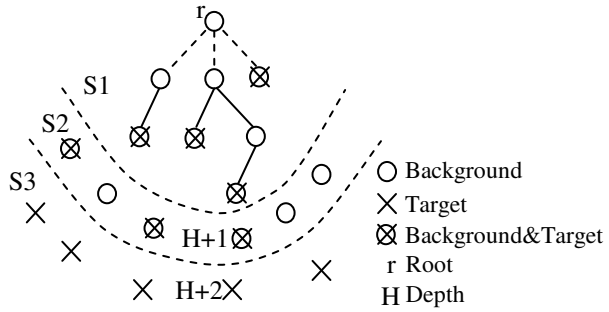

Figure 4: Locations of Keywords

In Figure 4, background keywords are shown as circles, and target keywords are displayed as crosses. Three areas, S1, S2, and S3, contain different types of keywords. S1 and S2 contain all background keywords. Because a document frequency threshold is applied when the concept hierarchy is developed, background keywords whose document frequency is less than the threshold are not present in the concept hierarchy. These keywords fall into area S2, and other keywords are placed in the hierarchy in area S1. Target keywords could fall into any of the three areas. Those that are not found in background keyword space (areas S1 and S2) form area S3, while others are in either S1 or S2. In Figure 4 the circle-cross symbols represent keywords that are in both background and target documents.

Given two keywords *X* and *Y*, there are two possibilities of there locations: *X* and *Y* are both in the concept 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 (the number of words in the path including *X* and *Y*) is assigned as the hierarchy distance between *X* and *Y*. In case the selected path includes the root, a penalty of 1 is added to the distance. In the second condition, *X* and *Y* are not both present in the concept hierarchy, so there is no real connection between them. In such cases, the hierarchy distance is defined as the summation of the distance between *X* and *r* and the distance between *Y* and *r*. The hierarchy distance between a keyword *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 occurs when two keywords both are in S3, which is 2(*H*+2). Table 1 summarizes the calculations.

Table 1: Hierarchy Distance Calculation

	S1	S2	S3
S1	Len(X-Y), or Len(X-Y)+1	Len(r-X)+ (H+1)	Len(r-X)+ (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)

4.4 Calculation of User-oriented Novelty. The semantic distance between two keywords *X* and *Y* is defined as the square root of the product of their occurrence distance and hierarchy distance. The reason to choose the square root of the product of the two components instead of average is that the semantic distance can be shortened by either distance component, not necessarily both.

5. Evaluation

We conducted a user study to investigate the performance of the user-oriented novelty measure in terms of identifying interesting (previously unknown and potentially useful) rules. The purpose is to compare the novelty scores generated by the system to the subjective ratings of the rule novelty and usefulness judgments made by human users.

5.1 Methodology. Eight PhD students with different majors were invited to participate in the user evaluation. The first step was collecting background documents. Participants were asked to provide a set of documents that they have collected for their research. Participants were also asked to enter their research interests in 2 to 4 phrases.

The second step was retrieving target documents. For each participant, we formulated a query from his/her research interests and search for articles from the Google Scholar search engine in PDF format (for easy full text downloading). The system then developed the background knowledge for each participant and mined association rules among noun phrases from the target documents. After that, novelty of the discovered rules was calculated, and normalized from 1 to 7.

The third step was evaluating the rules. For each participant, more than 10 thousand association rules were discovered. We randomly selected 9 rules at each novelty level (1 to 7) to create a sample of 63 rules for evaluation. The sample rules were presented to participants in a random order. Participants were asked to rate the novelty and usefulness of each rule in a 7-point scale (1 for the least and 7 for the most).

We evaluated the user-oriented novelty measure from two perspectives: novelty prediction accuracy and usefulness indication power, by comparing its scores to the actual user ratings. We compared our measure with other interestingness measures as well.

5.2 Results. The novelty prediction accuracy is defined as the correlation between an interestingness measure and the user novelty ratings. The correlations of four measures (S: support, C: confidence, WN: WordNet novelty, UN: user-oriented novelty) are presented in Table 3.

Table 3: Correlations for Novelty Prediction

P#	S	C	WN	UN
1	-0.09	0.18	0.38	0.68
2	-0.13	0.17	0.11	0.52
3	-0.39	0.10	0.43	0.56
4	0.02	-0.03	0.33	0.41
5	-0.27	0.27	0.11	0.28
6	0.01	0.05	0.19	0.23
7	-0.22	0.08	0.11	0.40
8	-0.25	-0.04	0.31	0.68
Mean	-0.165	0.098	0.246*	0.470**
Std.	0.143	0.107	0.132	0.169

* Better than S and C ($p < 0.01$)

** Better than S, C and WN ($p < 0.005$)

Since S and C are not designed for identifying novel rules, it may not be fair to compare them with WN and UN for novelty prediction. They were included as a baseline, and any novelty measure should perform better than them for novelty prediction. The result shows that UN and WN are better than the baseline as expected, and UN performs significantly better than WN for rule novelty prediction.

Similar analysis was done for usefulness indication power of different interestingness measures. Besides WN, seven objective measures were included because identifying useful rules is one of the goals of all interestingness measures. In [14], a total of 21 objective measures were studied, and they were classified into 7 groups according to their property similarities. We chose one measure from each group for comparison. The chosen objective measures were Support (*SP*), Odds ratio (α), Jaccard (ζ), Piatetsky-Shapiro's (*PS*), Gini Index (*G*), Klosgen (*K*) and Kappa (κ) [14]. The correlations are shown in Table 4.

Table 4: Correlations for Usefulness Indication

P#	*SP*	α	ζ	*PS*	*G*
1	-0.02	0.06	0.05	0.10	0.09
2	-0.03	-0.14	-0.16	0.01	-0.04
3	-0.20	0.08	0.04	-0.19	-0.13
4	0.24	0.12	-0.09	-0.14	-0.01
5	-0.01	0.19	0.08	0.16	0.23
6	-0.06	-0.12	-0.04	-0.08	0.13
7	-0.30	0.04	-0.03	0.14	0.14
8	-0.19	-0.11	-0.22	-0.22	-0.19
Mean	-0.071	0.015	-0.046	-0.028	0.027
Std.	0.164	0.123	0.106	0.151	0.144

Continued:

P#	*K*	*κ*	*WNN*	*UN*
1	0.07	0.12	0.23	0.65
2	-0.12	-0.16	0.05	0.12
3	-0.19	-0.05	0.12	0.21
4	0.02	0.14	0.37	0.47
5	-0.03	0.10	-0.01	0.17
6	0.02	-0.28	0.36	0.11
7	0.02	0.04	0.10	0.33
8	-0.19	-0.14	0.28	0.71
Mean	-0.050	-0.029	0.188*	0.346**
Std.	0.103	0.154	0.143	0.238

* Better than all objective measures ($p < 0.05$)
** Better than all other measures ($p < 0.05$)

Table 4 shows that UN generally has a higher correlation with the subjective rule usefulness ratings, which suggests that rules predicted to be novel by the system are also useful. UN also outperforms all seven objective measures and WN in terms of correlating with user usefulness ratings, i.e. indicating useful rules.

The analyses suggest that UN has a high correlation with the subjective rule novelty and usefulness ratings. By ranking the rules by the novelty score, the user can save significant time and effort when looking for interesting (novel and useful) patterns.

6. Conclusions

This paper presents a methodology for personalized knowledge discovery from text. It evaluates the interestingness of discovered association rules using the background knowledge developed from users' background documents. The experiment shows that the proposed measure has high novelty prediction accuracy and usefulness indication power. It outperforms other interestingness measure for identifying novel and useful rules.

We will continue our work in the following directions. (1) A comprehensive evaluation. A larger-scale user study is planned to evaluate the effects of other factors. (2) Provide users with the local context of rules. Some rules are short and difficult to comprehend without looking at the context in which the words are used. (3) Enhance the background document collecting function by recording the user's visit history and relevance feedback to automatically create the background documents.

7. References

[1] R. Agrawal and R. Srikant. Fast Algorithms for Mining Association Rules. *Proceedings of the VLDB Conference*, 1994.

[2] S. Basu, R. J. Mooney, K. V. Pasupuleti and J. Ghosh Evaluating the Novelty of Text-Mined Rules using Lexical Knowledge. *Proceedings of the Seventh ACM SIGKDD Conference*, pp. 233-238, San Francisco, CA, August 2001.

[3] E. Brill. Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging. *Computational Linguistics*, 1995.

[4] S. Brin, R. Motwani, J.D. Ullman and S. Tsur Dynamic Itemset Counting and Implication Rules for Market Basket Data. *Proceedings of the ACM SIGMOD Conference*, pp.255-264, 1997.

[5] R. Forsyth and R. Rada. *Adding an edge in Machine Learning: applications in expert systems and information retrieval*. Ellis Horwood Ltd, 1986.

[6] M. Klemettinen, H. Mannila, P. Ronkainen, H. Toivonen and A. I. Verkamo Finding Interesting Rules from Large Sets of Discovered Association Rules. *Proceedings of the CIKM Conference*, 1994.

[7] B. Liu, W. Hsu, L. F. Mun and H. Y. Lee Finding interesting patterns using user expectation. *IEEE Transactions on Knowledge and Data Engineering*, 11:817-832, 1999.

[8] B. Liu, Y. Ma and R. Lee Analyzing the interestingness of association rules from the temporal dimension. *IEEE International Conference on Data Mining*, Silicon Valley, CA, 2001.

[9] B. Padmanabhan and A. Tuzhilin A belief-driven method for discovering unexpected patterns. *Proceedings of the KDD Conference*, New York, 1998.

[10] B. Padmanabhan and A. Tuzhilin Unexpectedness as a measure of interestingness in knowledge discovery. *Decision Support Systems*, 27:303-318, Elsevier Science, 1999.

[11] G. Piatesky-Shapiro and C. Matheus The interestingness of deviations. KDD-94, 1994.

[12] M. Sanderson and B. Croft. Deriving concept hierarchies from text. *Proceedings of the SIGIR Conference*, pp. 206-213, 1999.

[13] A. Silberschatz and A. Tuzhilin. What Makes Patterns Interesting in Knowledge Discovery Systems. *IEEE Transactions on Knowledge and Data Engineering*, 8 (6), 1996.

[14] P. Tan, V. Kumar and J. Srivastava. Selecting the right objective measure for association analysis. *Information Systems*, 29(4):293-313, 2004.

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