Mining Knowledge from Text Collections Using Automatically Generated Metadata

John M. Pierre

Interwoven, Inc.
101 2nd Street, 4th Floor
San Francisco, CA 94105
jpierre@interwoven.com

Abstract. Data mining is typically applied to large databases of highly structured information in order to discover new knowledge. In businesses and institutions, the amount of information existing in repositories of text documents usually rivals or surpasses the amount found in relational databases. Though the amount of potentially valuable knowledge contained in document collections can be great, they are often difficult to analyze. Therefore, it is important to develop methods to efficiently discover knowledge embedded in these document repositories. In this paper we describe an approach for mining knowledge from text collections by applying data mining techniques to metadata records generated via automated text categorization. By controlling the set of metadata fields as well as the set of assigned categories we can customize the knowledge discovery task to address specific questions. As an example, we apply the approach to a large collection of product reviews and evaluate the performance of the knowledge discovery.

1 Introduction

Businesses and institutions often have a great deal of information technology infrastructure devoted to maintaining various repositories of text documents. These repositories include local and networked file systems, document management systems, email, discussion groups, and portals. Much of the collective knowledge in an organization is contained within these repositories of text documents. However, text repositories typically go untapped as a source for knowledge discovery. Instead documents are often used only once, and stored away for possible future reference. Users may have the ability to search for relevant documents and retrieve them based on a query. Analyzing large numbers of documents to find trends, patterns, and rules of thumb is usually not possible without a great deal of manual effort and expert knowledge. In order to recover and utilize the valuable knowledge contained in these document collections we must develop efficient and effective methods of analyzing large amounts of unstructured text.

In this paper we describe an approach to knowledge discovery in text collections. Our methodology uses automated text categorization to create structured metadata which can then be analyzed using traditional data mining techniques.

The result is a set of associations or rules involving a pre-defined set of concepts. By using this approach it is possible to efficiently analyze large document repositories and discover new knowledge. This approach combines aspects of computer science and cognitive science within a framework that solves a real world business problem.

Due to recent advances in computer science in the areas of text categorization[5] and text data mining[1] it is now possible to accurately and automatically analyze unstructured text documents. The use of efficient algorithms as well as the availability of ever powerful computers mean that large collections of text documents can be processed quickly, inexpensively, and with a minimum of human intervention. While computational methods still do not come close to matching the reasoning abilities of a human knowledge worker, they do allow tedious and time consuming tasks to be performed automatically and permit human users to concentrate on higher level problems. Even though a purely manual process may result in more accurate knowledge discovery, it is usually not feasible or cost effective to have a human spend the time and effort to read, understand, analyze, and extract comprehensive insights from a text collection containing tens of thousands of documents.

A crucial part of our approach is the use of *faceted metadata*[6], which is a specialized form of knowledge representation that draws from principles of cognitive science. Individual facets represent orthogonal conceptual dimensions. In our approach the set of facets as well as the set of possible concepts in each facet must be determined by a knowledge engineer. By controlling the set of facets as well as the available set of concepts in each facet we can customize the metadata database schema for targeted mining tasks. For example we can control the level of generality or specificity in the concepts, or constrain the mining to the discover relationships between concepts within specific facets of interest. This allows us to probe different aspects of the knowledge contained in the underlying document collection. In addition, the implied semantic relationship between different facets allows knowledge workers to more easily interpret the meaning of the mined rules and associations.

This approach provides a solution that allows a business or organization to repurpose large untapped text repositories to unlock the knowledge within. The ability to customize the knowledge discovery to answer specific questions is important to ensure actionable results. In our approach the metadata schema is not rigid and does not need to be determined before the documents are created or collected. This is in contrast to traditional database systems where the schema is fixed before data is gathered and therefore data mining is more constrained. Our flexible approach allows the schema to be adjusted to suit the knowledge discovery task, and takes advantage of the rich complexity of relationships between concepts that are inherent in a large document repository. This allows the user to formulate a specific hypothesis or question and to configure the system to test the hypothesis or answer the question based on the contents of the text collection.

In section 2 we discuss in detail the approach to creating suitable metadata as an input to the knowledge discovery process. In section 3 we provide an example where we applied the approach to a large collection of product reviews. Related work is discussed in section 4. We state our conclusions and suggestions for further study in section 5.

2 Methodology

The approach presented in this paper uses automated text categorization to assign faceted metadata records to text documents. These metadata records serve as a bridge between a corpus of free text documents and a highly structured database with a rigid schema. Statistical techniques and traditional data mining can then be applied to the set of structured metadata records to discover knowledge implicit in the underlying document collection. By choosing the metadata schema and the set of concepts in each facet we can control the knowledge discovery process.

In traditional data mining (or *Knowledge Discovery in Databases*) the database schema is rigid and is usually fixed before the data is even collected. In contrast a corpus of documents is inherently more flexible, and since the metadata schema can be changed at any time, the corpus can be re-purposed to address different kinds of questions.

Our basic approach is:

1. Gather a document collection that covers the domain of interest.
2. Segment documents into an appropriate set of transactions.
3. Construct a metadata schema with facets and concepts that suit the goal of the knowledge discovery task.
4. Train text classifiers to populate the metadata fields using machine learning techniques.
5. Apply automated text categorization to create a metadata database.
6. Use data mining to discover associations between concepts or derive rules.

In the rest of this section we describe these aspects in more detail.

2.1 Document Selection

Successful knowledge discovery requires a sufficient document collection. The document collection must span the domain of interest so that the concepts and associations between them are adequately represented. In addition the collection must provide enough transactions so that statistically significant rules can be mined. Furthermore, each document should be granular enough to provide crisp associations between a focused set of concepts.

2.2 Document Segmentation

In this framework, the issue of what constitutes a document is an important one. In our approach each document defines a "transaction" that creates the associations between concepts. In analogy with a traditional data mining example, each document is like a market basket and the concepts assigned in the metadata constitute the items in the basket. Given a collection of large documents it may be useful to further segment into chapters, paragraphs, passages, *etc.* to achieve the right level of granularity.

2.3 Metadata Schema

Metadata is often defined as "data about data" and when associated with documents is usually intended as an aid to searching, organizing, and summarizing in large collections. In this work we use metadata in a generalized way to gather structured information about free text documents. Faceted metadata is a form of knowledge representation akin to templates or frames and slots. Metadata facets represent orthogonal sets of concepts (for example "People", "Locations", "Dates"). By constraining the data mining to analyze the co-occurrences of concepts in particular facets we can exercise considerable control over the knowledge discovery process.

Most approaches to text mining use natural language processing or information extraction to select the set of keywords or phrases to be analyzed. This can lead to the "vocabulary problem" where differences in word usage such as synonyms, homonyms, or spelling errors can lead to spurious results. By establishing a fixed set of concepts in each facet we can control the vocabulary used in the rule mining phase.

2.4 Text Categorization

Creating metadata can be tedious and expensive, and it can lead to inconsistent results if done completely manually. Automated text categorization has become a practical way to create metadata records for large collections of documents. The main cost is in training and tuning a classifier. Classifiers are fairly easy to adjust, which allows the metadata schema to be changed according to the needs of the knowledge discovery project.

Text categorization is the assignment of relevant categories to documents, and a number of machine learning techniques have been developed to automate the process[5]. After training with a sufficient number of example documents, associations are made between the words in documents and the concepts in the categorization scheme. Several different classifiers can be trained to assign categories from conceptually distinct taxonomies to populate each metadata facet.

The accuracy of the text categorization is an issue. With sufficient training, automated classifiers can achieve close to human levels of accuracy, but there is always some inherent error rate. Since data mining usually relies on statistically

significant co-occurrences of concepts, inaccuracies in category assignments will be somewhat mitigated if they lead to spurious co-occurrences which are not statistically significant. However if classifier inaccuracies are biased (which is often the case) this will reduce the quality of the mined rules.

2.5 Data Mining

The data mining phase can be comprised of statistical techniques or algorithms to discover knowledge, rules, and relationships between concepts. Successful data mining leads to the discovery of facts which are non-trivial and previously unknown. In our approach this phase is more similar to traditional data mining in databases than to text mining because we operate on structured metadata records instead of directly on the free text itself. Since our concepts are typed by the metadata schema we can achieve greater semantic richness in our mined rules and associations than in traditional text mining [12].

In this paper we apply a common association rule induction technique to discover pair-wise relationships between concepts in different facets. However, more complicated learning algorithms could be applied to the metadata records to derive higher order rules.

3 Example

We apply our approach to a large collection of product reviews as an illustrative example. Individual text classifiers were trained to assign metadata for each facet, and data mining techniques were applied to discover associations between the concepts in different facets. The correct metadata records for each document as well as the expected set of associative relationships between concepts was known ahead of time and used as a basis for testing the quality of the text categorization and knowledge discovery phases.

3.1 Document Collection

To build a suitable document collection for experimentation, we spidered a web site containing product reviews for audio equipment[7]. The site arranged products into high-level product categories (*e.g* Speakers) and low-level subcategories (*e.g.* Main Speaker, Bookshelf Speakers, Subwoofers, *etc.*). Each product review included an overall rating as well as a free text summary expressing the reviewer's opinion of the product. We downloaded the most reviewed products in each category for a total of 47,923 individual product reviews.

We randomly split the document collection roughly in half to form a training set and test set. The first set of 24147 documents was used to train our text classifiers. Automated text categorization was performed on the second set of 23776 documents, and data mining was performed on the automatically created metadata.

3.2 Metadata Schema

The metadata was organized into four facets: Category, Subcategory, Products, and Rating. The metadata schema is shown in Table 1.

Table 1. Metadata Schema

Metadata Facet	Number of Concepts
Category	11
Subcategory	49
Products	1610
Rating	2

In our document collection there were 11 high-level product categories divided into 49 subcategories. A total of 1610 products were represented. Products were originally rated on a scale from 1 to 5 which we consolidated into two concepts, *GOOD* and *BAD*.

3.3 Classifier Training and Evaluation

We trained a separate text classifier for each of the four metadata facets. Separate Naive Bayes classifiers[8] were trained to assign product categories and subcategories. A simple Boolean classifier was constructed from the list of products derived from the pre-assigned metadata in the training set. The Boolean classifier checked if all tokens in a product name occurred in the document in order to score a match. We trained a Naive Bayes classifier to distinguish between "GOOD" and "BAD" product ratings. The text classifiers and their estimated performance are summarized in Table 2.

Table 2. Text Classifiers

Metadata Facet	Classifier	Precision	Recall
Category	Naive Bayes	0.79	0.79
Subcategory	Naive Bayes	0.54	0.54
Products	Boolean	0.31	0.18
Rating	Naive Bayes	0.91	0.91

Classifier performance was estimated using the standard micro-averaged precision and recall measures[9]. The Naive Bayes classifiers for Category, Subcategory, and Rating assigned only a single concept (and only a single concept was assumed to be correct), therefore precision and recall were equivalent.

The Boolean classifier for products was able to assign more than one product. For each review only one product was considered to be "correct" even though other (competing or complementary) products may have been mentioned in the summary or even if no products were explicitly mentioned at all. Therefore this strict interpretation of precision and recall somewhat underestimated the performance of the product classifier. A cursory inspection of the results indicated good performance of the classifier at finding products mentioned in the reviews.

3.4 Mining

To discover associations between concepts we applied the Apriori algorithm given in [10][11]. We limited our results to association rules of the form $A \rightarrow B$ with a single antecedent and consequent, each restricted to different specified facets.

Rules were selected according to thresholds on confidence and support. The support of a rule is defined by

$$Support(A) = (|A|/|T|) \times 100\%,$$

where $|A|$ is the number of transactions in which the set A occurs and $|T|$ is the total number of transactions.

The confidence of a rule is given by

$$Confidence(A \rightarrow B) = \frac{Support(A, B)}{Support(A)} \times 100\%.$$

We tested rules selected with support thresholds of 0.1% (corresponding to approximately 10 transactions), 0.05% (approx. 5 transactions), and 0.01% (approx. 1 transaction). In all cases we used a confidence threshold of 60%.

We mined four kinds of simple association rules:

$$Subcategory(A) \rightarrow Category(B)$$

$$Product(A) \rightarrow Category(B)$$

$$Product(A) \rightarrow Subcategory(B)$$

$$Product(A) \rightarrow Rating(B)$$

In the first case we found associations between instances of subcategories and high-level categories. In the second type of association rule product categories were inferred from product instances, and given products we inferred product subcategories in the third type. In the last type we found the association of individual products with a *GOOD* or *BAD* review.

3.5 Results

In Table 3 we show some selected examples for each rule type. Each rule type corresponds to the pair-wise association of concepts in two different facets. Though some of the associations are conceptually related between rules types, each was derived independently.

Table 3. Example Rules

Rule Type	Example
Subcategory(A) → Category(B):	
	A/V Receivers → Amplification
	DVD Players → Home Video
	Main Speaker → Speakers
Product(A) → Category(B):	
	Yamaha RX/V795 → Amplification
	Samsung DVD/611 → Home Video
	Paradigm Atom → Speakers
Product(A) → Subcategory(B):	
	Yamaha RX/V795 → A/V Receivers
	Samsung DVD/611 → DVD Players
	Paradigm Atom → Main Speaker
Product(A) → Rating(B):	
	Yamaha RX/V795 → GOOD
	Samsung DVD/611 → BAD
	Paradigm Atom → GOOD

3.6 Evaluation

We attempted to quantify the accuracy of mined association rules based on the ability of the system to re-derive a known set of associations between sets of concepts. To estimate the performance of our data mining tasks we have defined analogs of the standard precision and recall measures used in information retrieval and text categorization:

$$Precision = \frac{\#\,of\,correct\,rules\,mined}{\#\,of\,total\,rules\,mined}$$

$$Recall = \frac{\#\,of\,correct\,rules\,mined}{\#\,of\,rules\,known\,to\,be\,correct}$$

In order to compute these evaluation metrics we must have some prior knowledge of which rules are correct, which rules are incorrect, as well as the total number of possible rules. In more complicated mining tasks these values could be difficult or impossible to deduce, but using our constrained approach we are able to estimate them.

Since all products were known to be assigned to a specific product category and subcategory, we assumed that this defined the correct set of associations of the type $Product(A) \rightarrow Category(B)$ and $Product(A) \rightarrow Subcategory(B)$. To estimate the correct set of associations of the type $Product(A) \rightarrow Rating(B)$ we computed the average numerical score of the individual product reviews in our test collection and mapped them into the $GOOD$ and BAD categories. Each product category was divided into several subcategories which defined the set of correct associations for $Subcategory(A) \rightarrow Category(B)$.

The concept of collecting the "right answers" for evaluation of association rules is based on some key assumptions. When the association relationship in our control set is based on instances (*e.g.* products) assigned to classes (*e.g.* categories), the associations discovered from mining can be the result of other types of relations besides class membership. For example in our data the occurrence of particular kind of audio amplifier commonly used to test high-end speakers can result in an association that would be considered incorrect according to our evaluation but not necessarily a bad rule as far as useful knowledge discovery is concerned. Likewise a highly rated product might often appear in reviews for a product with a low rating as a good alternative. Another assumption is that a complete set of right answers can even be defined. Nevertheless similar assumptions are also made when evaluating information retrieval (relevance judgments) and text categorization (category assignments) systems, and this type of evaluation has proven useful as a basis for comparison and optimization despite its limitations.

Table 4. Rule Evaluation Results

Rule Type	Support Threshold	# of mined rules	Precision	Recall
Subcategory(A) → Category(B):				
	0.1%	16	1.0	0.33
	0.05%	20	1.0	0.41
	0.01%	26	1.0	0.53
Product(A) → Category(B):				
	0.1%	168	0.88	0.10
	0.05%	341	0.84	0.18
	0.01%	475	0.74	0.29
Product(A) → Subcategory(B):				
	0.1%	86	0.74	0.05
	0.05%	210	0.60	0.08
	0.01%	443	0.39	0.11
Product(A) → Rating(B):				
	0.1%	259	0.91	0.17
	0.05%	474	0.92	0.28
	0.01%	881	0.89	0.5

In Table 4 we present evaluation results for our mined association rules. For each rule type we show the total number of mined rules, precision, and recall at the specified minimum support thresholds of 0.1%, 0.05%, and 0.01%. In general lowering the support threshold resulted in decreased precision and increased recall. Even though precision decreased the number of "correct" mined rules generally increased substantially as the support threshold was lowered.

In most cases we were able to discover rules with high precision, but low recall. Though we have shown that recall can be increased at the expense of precision by adjusting the support threshold, it is also likely that we could achieve higher recall values given a larger document collection. The relatively high levels of precision achieved show that assumptions made for generating the sets of correct answers where generally justified.

The associations of the type $Product(A) \rightarrow Subcategory(B)$ resulted in lower performance values. This is probably because of the relatively poor performance of the classifier for subcategories. The fairly subtle distinctions between some subcategories (*e.g.* Amplifiers vs. Integrated Amplifiers), were difficult for our classifier to distinguish between. In these cases the classifier tended to be biased toward the subcategories that had the most documents in the training set, and this bias was reflected in the mined rules.

4 Related Work

Text mining and applications of data mining to structured data derived from text have been the subject of much research in recent years. Most text mining has used natural language processing to extract key terms and phrases directly from the documents[2][1].

Some approaches have used external knowledge as an enhancement. In [12], a focused set of terms was generated from the document collection and arranged into a hierarchical taxonomy to refine their mining tasks. Loh *et al.*[13] use automated categorization to assign a collection of pre-defined concepts to a corpus of documents. Statistical techniques were then applied to the sets of assigned concepts to find associative rules and concept distributions. However, their concepts required a significant amount of domain knowledge to construct via manual training and were not typed into facets.

In more recent work, machine learning techniques have been used to derive complex structured data from text to which data mining techniques such as rule induction can be applied. In [3], a knowledge base was constructed around a set of predefined conceptual entities (*i.e.* companies) and various web pages were analyzed using text categorization, information extraction, and wrappers to derive specific features for each entity. Traditional data mining was then applied to the derived knowledge base to discover various rules. Their approach, while very similar to what we've presented here, is more difficult to apply to a generic collection of unstructured documents since each of their input documents must be keyed to a specific entity and some amount of effort is required to develop customized wrappers for certain types of documents. In [4], information extraction was used to construct a database of structured records from a document corpus. Data mining was applied to the database to discover prediction and association rules. The accuracy of prediction rules was evaluated by measuring the average ability to predict each slot value based on all other slot values. Basu *et al.* [14] present a method to evaluate the quality of mined rules based on their "nov-

elty." A system for exploiting faceted metadata in a browsable user interface is described in [6].

5 Conclusions

By dynamically creating faceted metadata for a large collection of documents, we can construct a system for mining specific aspects of the knowledge implicit in the underlying corpus. Creating an appropriate set of text categorizers allows us to control the nature of the knowledge discovery and provides an automated system for deriving structured data from unstructured text. The main cost of the project is then shifted to collecting appropriate training data for the categorizers and defining facets and taxonomies. This type of system could be applied to a dynamic and growing document collection to monitor specific aspects of information that may change over time.

Our approach provides a practical solution for businesses and organizations that want to leverage and repurpose their document repositories in order to uncover useful knowledge. Since the metadata framework can be customized for specific domains, this approach could be applied to a wide variety of settings such as customer relationship management, human resources, competitive intelligence, message boards, intranets and the web. For example customer feedback email could be analyzed to determine relationships between user preferences, complaints and specific products, services or marketing campaigns. As another example, human resources documents such as resumes and performance reviews could be analyzed to find associations between people, departments and areas of expertise.

An interesting area of further study would be to explore the integration of this type of text based knowledge discovery with applications that feed on knowledge such as decision support systems, automated alert systems, agents, or expert systems. Another important area for further research is to understand how the presentation, filtering, and use of automatically discovered knowledge in the context of users' work processes leads to the most value.

References

1. M. A. Hearst. Untangling Text Data Mining. In *Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics*, 1999.
2. H. Ahonen and O. Heinonen. Applying Data Mining Techniques in Text Analysis. *Report C-1997-23*, University of Helsinki, Department of Computer Science, March 1997.
3. R. Ghani, R. Jones, D. Mladenic, K. Nigam, and S. Slattery. Data Mining on Symbolic Knowledge Extracted from the Web. In *Proceedings of the Sixth International Conference on Knowledge Discovery and Data Mining (KDD-2000) Workshop on Text Mining*, 29-36, 2000.
4. U. Nahm and R. Mooney. Text Mining with Information Extraction. In *Proceedings of the AAAI 2002 Spring Symposium on Mining Answers from Texts and Knowledge Bases*, 2002.

5. Y. Yang and X. Liu. A re-examination of text categorization methods. In *Proceedings of the 22nd Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, 42-49, 1999.
6. J. English, M. Hearst, R. Sinha, K. Swearingen, K.-P. Yee. Flexible Search and Navigation using Faceted Metadata. *Submitted for publication*, 2002.
7. AudioREVIEW.com
http://www.audioreview.com/
8. A. McCallum and K. Nigam. A Comparison of Event Models for Naive Bayes Text Classification. In *AAAI-98 Workshop on Learning for Text Categorization*, 1998.
9. D. Lewis. Evaluating Text Categorization. In *Proceedings of the Speech and Natural Language Workshop*, 312-318, 1991.
10. R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. I. Verkamo. Fast discovery of association rules. In U. Fayyad *et al.*, editors, *Advances in Knowledge Discovery and Data Mining*, 307-328. AAAI Press, 1996.
11. C. Borgelt. Apriori.
http://fuzzy.cs.uni-magdeburg.de/~borgelt/apriori/apriori.html
12. R. Feldman, M. Fresko, H. Hirsh, Y. Aumann, O. Liphstat, Y. Schler, M. Rajman. Knowledge Management: A Text Mining Approach. In *Proceedings of the 2nd International Conference on Practical Aspects of Knowledge Management (PAKM98)*, 29-30, 1998.
13. S. Loh, L. Wives, J. P. M. de Oliveira. Concept-based Knowledge Discovery in Texts Extracted from the Web. *SIGKDD Explorations*, 2(1): 29-39, 2000.
14. S. Basu, R. J. Mooney, K. V. Pasupuleti, and J. Ghosh. Evaluting the Novelty of Text-Mined Rules Using Lexical Knowledge. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2001)*, 233-238, 2001.
15. J. Han and Y. Fu. Discovery of Multiple-Level Association Rules from Large Databases. In *Proceedings of the 21st VLDB Conference*, 1995.