Text Mining - Knowledge extraction from unstructured textual data

Martin Rajman, Romaric Besançon
Artificial Intelligence Laboratory
Computer Science Dpt
Swiss Federal Institute of Technology

Abstract: In the general context of Knowledge Discovery, specific techniques, called Text Mining techniques, are necessary to extract information from unstructured textual data. The extracted information can then be used for the classification of the content of large textual bases. In this paper, we present two examples of information that can be automatically extracted from text collections: probabilistic associations of key-words and prototypical document instances. The Natural Language Processing (NLP) tools necessary for such extractions are also presented.

Key-words: text mining, knowledge discovery, natural language processing

1. Introduction

The general purpose of Knowledge Discovery is to "extract implicit, previously unknown, and potentially useful information from data" (Frawley, 1991). Due to the continuous growth of the volume of electronic data currently available, automated knowledge extraction techniques become always more necessary to valorize the huge amounts of data stored in the information systems. In addition, as the usual Data Mining techniques are essentially designed to operate on structured databases, specific techniques, called Text Mining techniques, have to be developed to process the important part of the available information that can be found in unstructured textual form.
Text Mining (TM) therefore corresponds to the extension of the more traditional Data Mining approach to unstructured textual data and is concerned with various tasks such as extraction of information implicitly contained in collections of documents or similarity-based structuring and visualisation of large sets of texts.
In sections 2 and 3, we present two examples of information extraction using different TM techniques: automated key-word association extraction, and prototypical document mining. In section 4, a discussion on the relation between these TM techniques and other Textual Data Analysis techniques is presented.

2. Mining for associations

In the case of an indexed document collection, the indexing structures (usually key-word sets) can be used as a basis for information extraction.
In such a framework, one possible goal is to extract significant key-word *associations*.

Let's consider a set of key-words $A = \{w_1, w_2, ..., w_m\}$ and a collection of indexed documents $T = \{t_1, t_2, ..., t_n\}$ (*i.e.* each t_i is associated with a subset of A denoted $t_i(A)$).
Let $W \subseteq A$ be a set of key-words, the set of all documents t in T such that $W \subseteq t(A)$ will be called the **covering set** for W and denoted $[W]$.
Any pair (W, w), where $W \subseteq A$ is a set of key-words and $w \in A \backslash W$, will be called an **association rule** (or simply an **association**), and denoted $(W \Rightarrow w)$.

Given an association rule $R : (W \Rightarrow w)$,

- $S(R,T) = |[W \cup \{w\}]|$ is called the **support** of R with respect to the collection T ($|X|$ denotes the size of the set X)
- $C(R,T) = \frac{|[W \cup \{w\}]|}{|[W]|}$ is called the **confidence** of R with respect to the collection T.

  Notice that $C(R,T)$ is an approximation (maximum likelihood estimate) of the conditional probability for a text of being indexed by the key-word w if it is already indexed by the key-word set W.

An association rule R generated from a collection of texts T is said to satisfy support and confidence constraints σ and γ if

$$S(R,T) \geq \sigma \text{ and } C(R,T) \geq \gamma$$

To simplify notations, $[W \cup \{w\}]$ will be often written $[Ww]$ and a rule $R : (W \Rightarrow w)$ satisfying given support and confidence constraints will be simply written as:

$$W \Rightarrow w \qquad S(R,T)/C(R,T)$$

Informally, for an association rule $(W \Rightarrow w)$, such σ/γ constraints can be interpreted as: there exists a significant number of documents (at least σ), for which being related to the topic characterised by the key-word set W implies (with a conditional probability estimated by γ) to be also related to the topic characterised by the key-word w.

As far as the actual association extraction is concerned, the common procedures are usually two-steps algorithms:

- generation of all the key-word sets with support at least equal to σ (*i.e.* all the key-word sets W such that $|[W]| \geq \sigma$). The generated key-word sets are called the **frequent sets** (or **σ-covers**);
- generation of all the association rules that can be derived from the produced frequent sets and that satisfy the confidence constraint γ.

The frequent sets are obtained by incremental algorithms that explore the possible key-word subsets, starting from the frequent singletons (*i.e.* the $\{w\}$ such as $|[\{w\}]| \geq \sigma\}$) and iteratively adding only those key-words that produce new frequent sets. This step is the most computationally expensive (exponential in the worst case) in the extraction procedure.
The associations derived from a frequent set W are then obtained by generating all the implications of the form $W \backslash \{w\} \Rightarrow w, (w \in W)$, and keeping only the ones satisfying the confidence constraint γ.
Some additional treatment (structural or statistical pruning, redundancy elimination) is usually added to the extraction procedure in order to reduce the number of generated associations.

Association-based Text Mining techniques have been explored by R. Feldman (Feldman, 1996) with the KDT (Knowledge Discovery in Texts) tool on the Reuter corpus. This corpus is a newswire collection containing about 22.000 articles manually indexed with 135 categories form the Economics domain by Reuters Ltd. and Carnegie Group Inc. in 1987.
For such a corpus, association extraction from the key-word sets allows to satisfy information needs expressed by queries. The result of the extraction process is a list of associations, ordered by support and confidence. An example of queries and associated results is given in Table 1.

Table 1: *Examples of associations*

query	:	'*find all associations between a set of countries including Iran and any person*'
result	:	[Iran, Nicaragua, Usa] $\Rightarrow$ Reagan 6/1.000
		...
query	:	'*find all associations between a set of topics including Gold and any country*'
result	:	[gold, copper] $\Rightarrow$ Canada 5/0.556
		[gold, silver] $\Rightarrow$ USA 18/0.692
		...

3. Mining for prototypical documents

Another direction of research for automated information extraction is to apply knowledge discovery techniques to the complete textual content of the documents (in a so-called "full text" approach as opposed to approaches only considering indexing key-words). However, our experiments on the Reuter corpus (Rajman, 1997) have shown that the extraction process does not produce any exploitable results when the standard association extraction techniques are directly applied on the words contained in the documents instead of operating on the already abstract concepts represented by the key-words. Among the extracted associations, some only indicate the presence of domain-dependent compounds ($\{wall\} \Rightarrow street$, $\{government\ prime\} \Rightarrow minister$, $\{treasury\ secretary\ james\} \Rightarrow baker$), while others are simply uninterpretable ($\{dollars\ shares\ exchange\ total\ commission\ stake\} \Rightarrow securities$, $\{million\ april\ management\ lead\ issues\ underwriting\ denominations\} \Rightarrow selling$).

A different approach is therefore necessary when full text is considered: prototypical document extraction. A *prototypical document* is informally defined as a document corresponding to an information that occurs in a repetitive fashion in the document collection, *i.e.* a document representing a class of similar documents in the textual base.
The extraction techniques operating in this framework still use the notion of frequent sets, but additional Natural Language (NL) techniques are used to preprocess the data, and identify more significant linguistic entities (*terms*) for the frequent set extraction process.
More precisely, the NL preprocessing, realized in collaboration with R. Feldman's team at Bar Ilan University, was decomposed into two steps: *Part-of-Speech tagging* and *term extraction.*

Part-of-Speech tagging
This process automatically identifies the morpho-syntactic categories (noun, verb, adjective, ...) of words in the documents. Such a tagging allows to filter non-significant words on the basis of their morpho-syntactic category. In our experiments, we used a rule-based tagger designed by E. Brill (Brill, 1992), and restricted the extraction process to operate only on nouns, adjectives and verbs.

Term extraction
This process aims at the identification of the domain-dependent compounds. It allows the mining process to focus on more meaningful co-occurrences, and can be decomposed into:

- term candidates identification (on the basis of structural linguistic

information; in our case, morpho-syntactic patterns such as *Noun Noun*, *Noun of Noun*,...);

- term candidates filtering (based on statistical relevance scoring (Daille, 1994)).

For example, the already mentioned sequence of words (found by association extraction on full text) *(Treasury, Secretary, James, Baker)* was tagged as *Treasury/Noun Secretary/Noun James/Noun Baker/Noun* and subsequently identified as one single term *Treasury_Secretary_James_Baker/Noun*.

Mining process

On the basis of the terms resulting from the NL preprocessing step, an algorithm similar to the one described in the previous section was used to extract frequent term sets from the document collection.
The extracted frequent sets were then submitted to several additional treatments in order to determine the prototypical documents:

- To reduce information redundancy, a clustering of the frequent term sets was performed, on the basis of a similarity measure derived from the number of common terms in the sets. The resulting clusters were represented by the union of their constitutive term sets.

- To limit the possible meaning shifts due to variations in the word ordering, the clusters were further split into sets of *term sequences* associated with paragraphs boundaries in the original documents.

An example of the treatments on the Reuter corpus is given in Figure 1.

Figure 1: *The frequent sets processing*

Some frequent term sets (with their frequency) extracted from the Reuter Corpus for a support of 80:
{due available management priced issuing denominations payment_date} 87
{due management issuing denominations luxembourg payment_date} 81
{due management priced issuing combined paying underwriting} 80
{due management selling priced issuing listed} 81
{due priced issuing combined denominations payment_date} 80
{management issuing combined underwriting payment_date} 80
(...)

Resulting cluster:
{due available management priced issuing combined denominations listed underwriting luxembourg payment_date paying} 45

Most frequent sequential decomposition:
(issuing due paying priced) (available denominations listed luxembourg) (payment_date) (management underwriting combined) 41

Prototypical documents are then all the documents (or document parts) that instantiate any of the extracted sequential decompositions of the frequent term set. An example of a prototypical document instantiating the above mentioned decomposition is shown in Figure 2.

Figure 2: *A prototypical document*

<DOC2088>
Nissan_Motor_Co_Ltd "NSAN.T" is **issuing** a 35_billion_yen eurobond **due** March_25 1992 **paying** 5-1/8_percent and **priced** at 103-3/8, Nikko_Securities_Co (Europe) Ltd said.
The non-callable_issue is **available** in **denominations** of one_million Yen and will be **listed** in **Luxembourg**.
The **payment_date** is March_25.
The selling_concession is 1-1/4_percent while **management** and **underwriting combined** pays 5/8_percent.
Nikko said it was still completing the syndicate.

By definition, these prototypical documents are representative of classes of repetitive document structures in the collection of texts. Their main advantage is to provide a usable interpretation scheme for the information extracted from the document collection in the form of frequent term sets, and, as such, they constitute good candidates for a partial synthesis of the information content hidden in a textual base.

4. Related Work

Several other domains concerned with Textual Data Processing (such as Textual Data Analysis or Content Analysis) can provide interesting insights on the techniques presented in this paper.
The problem of frequent set extraction could be for instance partially related to the identification of co-occurrent words (Lafon, 1981), repeated segments (Salem, 1987), or quasi-segments (Becue, 1993), often considered in the domain of Textual Data Analysis. The main difference here is that the Text Mining techniques rely on the use of frequencies of *sets of words* instead of considering co-frequencies of *pairs*.
As far as more sophisticated information extraction is concerned, methods used in Textual Data Analysis (Lebart, 1998) usually rely on a cluster analysis based on the the chi-square distance between the lexical profiles. For each of the resulting clusters of documents, *characteristic words* (Lafon, 1980) (i.e. words with a frequency in the cluster significantly higher

than the one expected according to a predefined probabilistic model) are then extracted. Each of the cluster is then represented by a *characteristic document* which is the document in the cluster that contains the most characteristic words.
The differences between such approaches and prototypical document extraction as described in this paper are essentially of two kinds: (1) prototypical document extraction integrates a more substantial amount of explicit linguistic knowledge, in particular in the preprocessing phase, where morpho-syntactic patterns are used for the extraction of indexing terms; (2) the aims underlying the two methods are in fact quite different: documents characteristic for a cluster identify the information content that is the more discriminant for the cluster relatively to the rest of the document collection. On the opposite, prototypical documents tend to identify repetitive patterns of texts particularly frequent in the document collection, and that will serve to structure its informational content.
The two approaches therefore appear to be rather complementary in the sense that prototypical documents could be thought as kinds of linguistic frames in which the informational content (as identified by the characteristic documents) could be preferentially expressed.

In addition, in order to allow better representativity, a more generic representation could be achieved by using *name entity tagging*, a semantic tagging that allows to identify and generalise certain elements of a sentence. Such a tagging could lead to representations where the variable parts of the prototypical documents would be replaced by concepts.
For instance, on the basis of the results of name entity tagging applied to the document given in Figure 2 (these results were produced by the Alembic tool and provided by Christopher Clifton, from the MITRE NLP group (MITRE, 1997)), the associated document class could be represented by the generic prototypical document presented in Figure 3.

Figure 3: *A generic prototypical document*

<*ORGANIZATION*> is **issuing** a <*NUMBER*> yen eurobond **due** <*DATE*> **paying** <*NUMBER*> percent and **priced** at <*NUMBER*> , <*ORGANIZATION*> (<*LOCATION*>) said.
The non-callable issue is **available** in **denominations** of <*NUMBER*> and will be **listed** in <*LOCATION*>.
The **payment date** is <*DATE*>.
The selling concession is <*NUMBER*> percent while **management** and **underwriting combined** pays <*NUMBER*> percent.
<*ORGANIZATION*> said it was still completing the syndicate.

5. Conclusion

We have presented two examples of Text Mining tasks for the extraction of information from collections of textual data: an association extraction method operating on indexed documents, and a prototypical document extraction algorithm that can be applied on plain documents (full text approach). For both tasks, preliminary results have been obtained. Further research will be carried out to explore the use of prototypical documents for the automated synthesis of the information content of document classes in large collections of textual data.

References

Becue M., Peiro R. (1993). Les quasi-segments pour une classification automatique des réponses ouvertes, in *Actes des secondes journées internationales d'analyse des données textuelles*, (Montpellier), ENST, Paris, 310-325.

Brill E. (1992). A simple Rule-Based Part-of-Speech Tagger, in *Proc. of the 3rd Conf. on Applied Natural Language Processing.*

Daille B.(1994). Study and Implementation of Combined Techniques for Automatic Extraction of Terminology, in *Proc. of the 32nd Annual Meeting of the Association for Computational Linguistics.*

Feldman R., Dagan I. and Kloegsen W. (1996). Efficient Algorithm for Mining and Manipulating Associations in Texts, in *Proc. of the* 13th *European Meeting on Cybernetics and Research.*

Frawley W.J., Piatetsky-Shapiro G., and Matheus C.J. (1991). Knowlegde Discovery in Databases : An Overview, in *Knowledge Discovery in Databases*, MIT Press, pages 1–27.

Lafon P. (1980). Sur la variabilité de la fréquence des formes dans un corpus, *Mots*, 1, 127–165.

Lafon P. (1981). *Dépouillements et statistiques en lexicométrie*, Slatkine-Champion, 1984, Paris.

Lebart L., Salem A., Berry L. (1998). *Exploring Textual Data*, Kluwer Academic Publishers, Dordrecht.

MITRE NLP Group (1997). Alembic Language Processing for Intelligence Applications. At URL :
http://www.mitre.org/resources/centers/advanced_info/g04h/nl-index.html

Rajman M. and Besançon R. (1997). A Lattice Based Algorithm for Text Mining. Technical Report TR-LIA-LN1/97, Swiss Federal Institute of Technology.

Salem A. (1987). *Pratique des segments répétés, Essai de statistique textuelle*, Klincksieck, Paris.