Transactions on Information and Communications Technologies vol 29, © 2003 WIT Press, www.witpress.com, ISSN 1743-3517

Preprocessing method and similarity measures in clustering-based text mining: a preliminary study

S. Iiritano[1,2], M. Ruffolo[1,3] & P. Rullo[1,4]
[1] *Exeura s.r.l.*
[2] *DEIS—Dipartimento di Elettronica, Informatica e Sistemistica*
[3] *ICAR-CNR—Istituto di CAlcolo e Reti ad alte prestazioni del Consiglio Nazionale delle Ricerche*
[4] *Dipartimento di Matematica*
Università della Calabria, 87036 Rende (CS), Italy

Abstract

Knowledge Discovery in Text (KDT) has emerged as a challenging application due to the large amount of textual documents available from heterogeneous sources. An approach to knowledge discovery in text is based on clustering techniques in which the quality of results strongly depends on features extracted from documents and on similarity coefficients defined on them. In this work we present a framework for textual document preprocessing useful for the extraction of relevant features (i.e. lemma and word). Moreover, we define two similarity coefficients to measure "semantic" similarity (i.e. similarity by high-level contents, as they are intended by a human reader) among documents.

1 Introduction

Managing the huge amount of textual documents available on the web and on the intranets has become an important problem of knowledge management. For example, content management, one basic functionality of the knowledge management tools, deals with the documents categorization problem where groups of documents with similar contents are created.

Textual documents collections can be viewed as sources of unstructured data in which knowledge discovery is possible by using Knowledge Discovery in Text

Transactions on Information and Communications Technologies vol 29, © 2003 WIT Press, www.witpress.com, ISSN 1743-3517

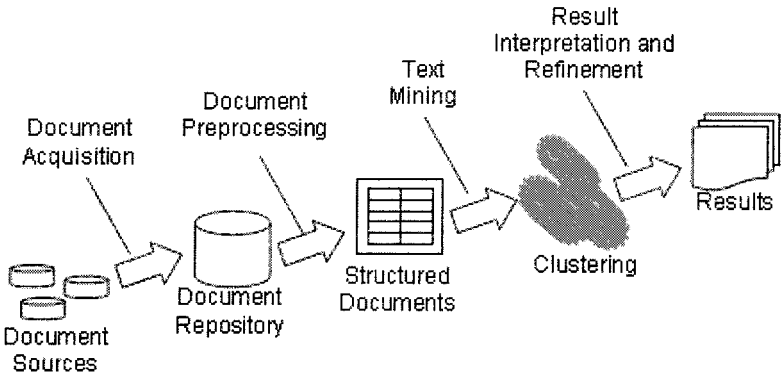


Figure 1: The Knowledge Discovery in Text process.

(KDT). Like Knowledge Discovery in Database (KDD), KDT is an interactive and iterative process, where the fundamental activity is the text mining, focused on the exploration of textual documents to discover new and interesting patterns within them [6].

A particular text mining approach is based on clustering techniques used to classify documents, similar by contents, in non-overlapped groups (clusters) on the basis of suitable similarity coefficients. The problem of text classification has been widely studied and many system have been developed [9, 23, 25, 26]. Several clustering algorithms [5, 11] as well as a wide variety of ad hoc similarity coefficients [2, 12, 24] were defined. However, the current clustering-based approaches to KDT exhibit limitations due to the poor investigation of the impact of the textual document preprocessing and of the type of extracted features on the existing similarity coefficient performances.

In this paper we propose a textual documents preprocessing framework for the extraction of features (i.e. words and lemma) usable to define two ad hoc similarity coefficients. These coefficients measure the "semantic" similarity of textual documents (i.e. similarity by high-level contents, as they are intended by a human reader).

2 Background on knowledge discovery in text

Figure 1 shows the structure of a KDT process [7].

The process is composed by the following phases:

- *Document Acquisition Phase*: a collection of documents (corpus) is collected in a repository. Generally, documents, coming from various sources, are of different types;
- *Document Preprocessing*: relevant features of documents, that describe their content, are extracted; finally, a suitable representation based on these features is created;

Transactions on Information and Communications Technologies vol 29, © 2003 WIT Press, www.witpress.com, ISSN 1743-3517

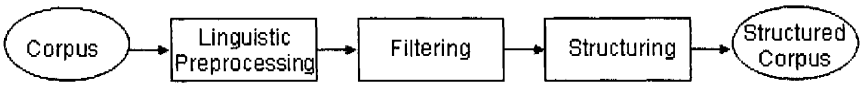


Figure 2: Document preprocessing framework.

- *Text Mining*: documents are treated using a text mining method. In literature several approaches has been proposed;
- *Results Interpretation and Refinement*: results of the text mining phase are submitted to performance evaluation using quality indexes [2] and to the interpretation and refinement of a human operator.

3 A framework for document preprocessing

In this section we show the proposed textual document preprocessing framework. Preprocessing is based on the following concepts:

- *Textual Document.* $\Delta = \{\varphi_1, \ldots, \varphi_\lambda\}$ is a sequence of λ features. In this work we assume that a feature can be a word or a lemma of a given language;
- *Corpus.* $\Omega = \{\Delta_1, \ldots, \Delta_n\}$ is a set of n textual documents;
- *Vocabulary.* $V = \{\varphi_1, \ldots, \varphi_m\}$ is the set of m distinct features occurring in a textual document;
- *Structured Vocabulary.* Given a textual document Δ and its vocabulary V the structured vocabulary is a set of couple $\Psi = \{(\varphi_j, f_r(\varphi_j))\}$ with $j = 1, \ldots, m$, where $f_r(\varphi_j)$ represents the *relative frequency* of the feature φ_j (i.e. the number of times that φ_j occurs in a textual document normalized w.r.t. the document length λ).

As shown in figure 2 the document preprocessing framework is based on three main activities:

- *Linguistic Pre-Processing*: textual documents are submitted to the operations of *POS-Tagging* and *lemmatization* [15]. These operations produce a tagged file that contains all the words of the document and, for each word, its lemma (obtained reducing the syntactical variants of the words to a unique form eliminating plurals, tenses, gerund forms [17]), and the part of speech represented by the word (i.e. article, verb, preposition, substantive, etc.). Linguistic preprocessing aims to guarantee a "machine readable" representation of the document contents;
- *Filtering*: each document in the corpus is submitted to *Explicit* and *Implicit Stop-features Removal*. In particular:
  - explicit stop-features occur frequently in a textual document and have a very low informative content (i.e. articles, prepositions, conjunc-

Transactions on Information and Communications Technologies vol 29, © 2003 WIT Press, www.witpress.com, ISSN 1743-3517

tions, adverbs, common verbs and adjectives, recognized through the *POS-Tagging*) [27];
  - implicit stop-features occur uniformly in the corpus (i.e. they have the same frequency in each document).
- *Structuring*: this step is performed to do a representation of the documents suitable to define similarity coefficients usable in clustering-based text mining approach. Given a filtered document its structured vocabulary is obtained.

For a given corpus the output produced by the textual document preprocessing framework is a set of structured vocabularies, one for each input document, called *Structured Corpus*.

In the next section we use the representation of the documents in term of structured vocabulary, obtained from the presented framework, to define two similarity coefficients.

4 Similarity measures definition

Performances of clustering-based text mining techniques are strongly affected by features chosen to represent the document and similarity coefficients defined on the extracted features.

Using the document representation (in terms of structured vocabulary), obtained through the preprocessing phase described above, each document can be viewed as a vector in the space of its features. Starting from this consideration two semantic similarity measures can be defined.

Let $W = V_{\Delta_i} \cap V_{\Delta_j} = \{\varphi_1, \ldots, \varphi_\mu\}$, be the set containing common features between documents, with $i, j = 1, \ldots, m$, and:

$$T = V_{\Delta_i} - (V_{\Delta_i} \cap V_{\Delta_j}) = \{t_1, \ldots, t_\tau\}$$
$$R = V_{\Delta_j} - (V_{\Delta_i} \cap V_{\Delta_j}) = \{r_1, \ldots, r_\rho\}$$

two sets containing non common features.

Considering only common features, contained in W, we can define the first similarity measures (SSM1):

$$SSM1(\Delta_i, \Delta_j) = \Pi_{ij} \cdot (1 - \Theta_{ij})$$

where

$$\Pi_{ij} = \frac{\sum_{k=1}^{\mu} f_{r\Delta_i}(\varphi_k) + \sum_{k=1}^{\mu} f_{r\Delta_j}(\varphi_k)}{2},$$

$$\Theta_{ij} = \sum_{k=1}^{\mu} \frac{|f_{r\Delta_i}(\varphi_k) - f_{r\Delta_j}(\varphi_k)|}{\mu}.$$

Note that $SSM1(\Delta_i, \Delta_j) \in [0, 1]$. In fact:

- if $W = V_{\Delta_i} \cap V_{\Delta_j} = \emptyset$ then $\Pi_{ij} = 0$, $\Theta_{ij} = 0$ and $SSM1(\Delta_i, \Delta_j) = 0$;
- if $W \equiv V_{\Delta_i} \equiv V_{\Delta_j} \neq \emptyset$ and $f_{r\Delta_i}(\varphi_k) = f_{r\Delta_j}(\varphi_k)$ for any $k = 1, \ldots, \mu$ then $\Pi_{ij} = 1$, $\Theta_{ij} = 0$ and $SSM1(\Delta_i, \Delta_j) = 1$;

Transactions on Information and Communications Technologies vol 29, © 2003 WIT Press, www.witpress.com, ISSN 1743-3517

Moreover, considering both common and non common features contained in T and R, we define the second semantic similarity measure (SSM2):

$$SSM2(\Delta_i, \Delta_j) = \Psi_{ij} - \Gamma_{ij}.$$

Where:

$$\Psi_{ij} = \Pi_{ij} - \Theta_{ij}$$

and

$$\Gamma_{ij} = \frac{\sum_{k=1}^{\tau} f_{r\Delta_i}(t_k) + \sum_{k=1}^{\rho} f_{r\Delta_i}(r_k)}{2}.$$

Note that $SSM2(\Delta_i, \Delta_j) \in [-1, 1]$. In fact:

- if $W = V_{\Delta_i} \cap V_{\Delta_j} = \emptyset$ then $\Pi_{ij} = 0$, $\Theta_{ij} = 0$, $\Gamma_{ij} = 1$ and $SSM2(\Delta_i, \Delta_j) = -1$;
- if $W \equiv V_{\Delta_i} \equiv V_{\Delta_j} \neq \emptyset$ and $f_{r\Delta_i}(\varphi_k) = f_{r\Delta_j}(\varphi_k)$ for any $k = 1, \ldots, \mu$ then $\Pi_{ij} = 1$, $\Theta_{ij} = 0$, $\Gamma_{ij} = 0$ and $SSM2(\Delta_i, \Delta_j) = 1$;

Both Similarity measures are based on the idea that two documents are identical when are formed by the same features distributed in the same way, but, in general, the values assumed by the two similarity measures have different behavior, in particular, in SSM1 the values are influenced only by the common features, whereas, in SSM2 the values are obtained as combination of common and non common features.

5 Conclusions and future works

In this preliminary work we have defined a framework for textual document preprocessing that performs a linguistic and a structural preprocessing and extract two type of features (lemma and word) from textual documents. The central idea of our work is that a document can be represented as a vector in the space of its features starting from the structured vocabularies obtained using the proposed framework. Moreover, on the base of this representation, we have defined two measures able to quantify the semantic similarity between two textual documents.

As future work we intent to implement the entire KDT process in a tool. In particular, the text mining phase based on clustering will be implemented using measures defined in this paper. The obtained tool will be tested using a corpus of textual documents concerning different topics.

References

[1] H. Ahonen, O. Heinonen, M. Klemettinen and A. I. Verkamo. Applying Data Mining Techniques in Text Analysis. Technical report, University of Helsinki.

[2] R. Baeza-Yates and B. Riberio-Neto. *Modern Information Retrieval.* Addison-Wesley, 1999.

[3] S. Chakrabarti. Data Mining for Hypertext: A Tutorial Survey. *ACM SIGKDD Explorations*, 1(2):1–11, January 2000.

Transactions on Information and Communications Technologies vol 29, © 2003 WIT Press, www.witpress.com, ISSN 1743-3517

[4] I. Dhillon and D. Modha. Concept Decompositions for Large Sparse Data Using Clustering. *Machine Learning*, 42:143–175, 2001.
[5] B. Everitt. *Cluster Analysis* Sage Publication, 1984.
[6] R. Feldman and I. Dagan. Knowledge Discovery in Textual Databases(KDT). *KDD'95: Knowledge Discovery and Data Mining, 112–117*, 1995
[7] R. Feldman, M. Fresko, H. Hirsh, Y. Aumann, O. Liphstat, Y. Schler and M. Rajman. Knowledge Management: a Text Mining Approach. *Proc. Of the Second Int. Conf. On Practical Aspects of Knowledge Management (PAKM98)*, 1998.
[8] W. J. Frawley, G. Piatesky-Shapiro and C. Matheus. Knowledge Discovery in Databases: an Overview *AI Magazine*, pp. 57-70, 1992.
[9] M. Grobelnik and D. Mladenic. Efficient Text Categorization. Text Mining Workshop on ECML-98.
[10] M. Halkidi, Y. Batistakis and M. Vazirgiannis. On Clustering Validation Techniques. Technical report, Department of Informatics, Athens University of Economics & Business.
[11] J. A. Hartigan. *Clustering Algorithms*. John Wiley and Sons, 1975.
[12] L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data* . John Wiley and Sons, 1989.
[13] B. Lent, R. Agrawal and R. Skirant. *Discovering Trends in Text Databases*. KDD'97: Knowledge Discovery and Data Mining 227–230, 1997.
[14] M. Lenz. *Managing the Knowledge Contained in Technical Documents*. PAKM'98: Practical Aspects of Knowledge Management, 1998.
[15] L. Marquez. *Part-Of-Speech Tagging: A Machine Learning Approach Based on Decision Tree*. Technical Report of the Technical University of Catalunya, 2000.
[16] A. McCallum and K. Nigam. A Comparison of Event Models for Naive Bayes Text Classification. In *In AAAI-98 Workshop on Learning for Text Categorization*, 1998.
[17] A. Miller Wordnet In Available at http://www.cogsci.princeton.edu/ wn/, Princeton University, 2002.
[18] M. F. Moens. *Automatic Indexing and Abstracting of Document Texts*. Kluwer Academic Publishers, 2000.
[19] M. Rajaman and R. Besancon. Text Mining: Natural Language Techniques and Text Mining Applications. *IFIP '97*, 1997.
[20] M. Ruffolo and S. Iiritano Managing the Knowledge container in Electronic Documents: A Clustering Method for Text Mining. *Theory and Application of Knowledge Management, DEXA Workshop Proceedings*, 2001.
[21] H. Schutze and C. Silverstein. Projections for Efficient Document Clustering. *SIGIR '97: Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval 74–81*, 1997.
[22] S. Soderland. Learning to Extract Text-Based Information from the World Wide Web. *KDD'97: Knowledge Discovery and Data Mining 251–254*, 1997.

Transactions on Information and Communications Technologies vol 29, © 2003 WIT Press, www.witpress.com, ISSN 1743-3517

[23] M. Steinbach, G. Karypis and V. Kumar. A Comparison of Document Clustering Techniques. *ACM-SIGKDD Workshop on Text Mining*, 2000.

[24] A. Strehl, J. Ghosh and R. Mooney. Impact of Similarity Measures on Web-Page Clustering. In *Proc. of AAAI Workshop on Artificial Intelligence for Web Search*, pp. 58–64, July 2000.

[25] O. Zamir and O. Etzioni. Web Document Clustering: A Feasibility Demonstration. In *Proc. of the 21st Annual International ACM SIGIR*, pp. 46–54, 1998.

[26] O. Zamir, O. Etzioni, O. Madani and R. M. Karp. Fast and Intuitive Clustering of Web Documents. *AAAI*, 1997.

[27] Y. Yang. Using Corpus Statistics to Remove Redundant Words in Text Categorization. *JASIS '96: Journal of the American Society for Information Science 357–369*, 1996.

[28] Y. Yang and J. O. Pedersen A Comparative Study on Feature Selection in Text Categorization. *International Conference on Machine Learning 412–420*, 1997.

[29] Y. Yang. An Evaluation of Statistical Approach to Text Categorization. *Information Retrieval 69–90*, 1999.