PROJECT MUSE®

Text mining and information retrieval

Canadian Journal of Information and Library Science, Volume 35,
Number 3, September/septembre 2011, pp. 223-227 (Article)

Published by University of Toronto Press

DOI: <https://doi.org/10.1353/ils.2011.0020>

➔ *For additional information about this article*

<https://muse.jhu.edu/article/450026>

Text mining and information retrieval

Introduction

In the past several years, many projects have been initiated to digitize and make available in digital format the information assets of organizations and branches of knowledge. By facilitating access to digital resources and increasing their quality of encoding and metadata, these projects have also motivated development of more effective techniques for research and analysis of textual information. Thus, with the increasing number of digital resources, techniques and strategies have been proposed to assist more effectively in the research and analysis of textual material—on the web or within documentation generated by organizations.

The interest in text-mining techniques for analysis and management of digital information lies in this area. Moreover, this domain is an important research area, in which concepts and techniques are derived mainly from work in data mining, artificial intelligence, and machine learning. However, conventional text mining (automatic clustering, automatic categorization, named entity recognition, etc.) also integrates concepts and techniques that have emerged from the field of computational linguistics. In this context, text mining combines analytical techniques from numerical and symbolic approaches based on linguistic textual data processing.

Recently, several researchers (Feldman and Sanger 2006; Ibekwe-SanJuan 2007; Srivastava and Sahami 2009; Weiss et al. 2005) have stressed the importance of exploring the appropriateness of applying techniques of text mining to information retrieval from the perspective of information science. In this effort, some online search-engine prototypes based on automatic text clustering have been developed (Carrot search, Clusty, etc.). Preliminary results of these prototypes have proven to be most relevant. However, few studies have actually explored and rigorously evaluated the appropriateness of using text-mining techniques in the context of information retrieval.

* * *

This special issue of the *Canadian Journal of Information and Library Science* presents the results of recent research into text-mining techniques using an information retrieval perspective. The four articles comprising this volume have been selected from sixteen articles submitted for review. The evaluation was done by leading specialists in these fields.

The paper by Andreani and his colleagues, "Normalisation des entités nommées : allier règles déclaratives, ressources endogènes et processus centré sur l'utilisateur," presents techniques for standardization of named entities encountered in documents processed by TecKnowMetrix. Named entities are expressions that denote individuals or "unique entities" such as geographic locations, names of persons, organizations or products, dates, etc. The named entities considered by these authors are primarily names of organizations that are pervasive in their corpus of patents, scientific publications, and techno-economic press articles. Their concern is to reduce all variants of the same named entity, such as Mitsubishi, Mitsubishi KK, and Mitsubishi Corp., to a single canonical form. This standardization is essential to ensure the reliable identification of all information relating to a particular entity. And it cannot benefit from linguistic resources such as lexical or terminological dictionaries, since it applies to proper names that have little to no regulation.

The normalization techniques used include (1) the use of specific vocabularies that identify the presence, the type, and sometimes the nationality of an organization, and a dictionary of country names, (2) the decomposition of complex names into their constituent base, (3) the gradual rewriting of organization names to conform to a standardized version, (4) a match of the resulting form to other, previously standardized forms, and (5) the identification of common sub-sequences of a given expression within the corpus. Their interactive approach is based on rules and on validation by a user.

Results from an expert's evaluation are presented, distinguishing between cases where standardization is correct, partial, or incorrect (with noise or silence). Exact normalizations (depending on the type of publication) represent an average of 84% for names of organizations, 83.8% for type ("academic" or "business"), and 62.4% for organizations' country of origin. Results are lower when combining two or three criteria. More specifically, the evaluation highlights the fact that 86.4% of named entities have had to undergo standardization before reaching the canonical expected form, indicating that the procedure is essential for efficient retrieval.

The article illustrates the usefulness of text-mining techniques (here, based significantly on natural language processing) for information retrieval.

The aim of “Bilingual document clustering: Evaluating cognates as features” by Denicia-Carral and his colleagues is to cluster documents written in two different languages, so that documents about the same topic are grouped in the same class, regardless of their language of expression. Clustering is usually done by using “features” for each document and grouping documents that share enough common features. The challenge is always to choose the most discriminating features. In this paper, the main features used to perform the clustering are cognates: words that are spelled identically or almost identically in both languages, such as *construction* or *céréale/cereal* in French and English. The idea is that cognates are very easy to find (very little language processing is required) and sufficiently reliable as a unifying feature. (Of course, some are false cognates, such as *librairie* and *library*, but their approach relies on the relative rarity of this phenomenon to minimize problems.)

Thus, the objective is achieved independently of external linguistic resources, and is in fact largely language independent—but the authors note the need for languages to be relatively close (same alphabet, many common roots, etc.) for this technique to be effective.

Two different methods of cognate extraction were used: extraction of similar pairs (based on a calculation of graph similarity), and extraction of similar pairs whose context is also similar, containing pairs of named entities that are identical. Also, two different methods of clustering were used (Direct and Star).

The evaluation was performed on a bilingual English-Spanish corpus. Two baselines were defined, which correspond to other approaches to the same problem: a clustering set created after translating documents from one language to another, and a clustering based solely on the named entities. Results exceed the performance of these baseline references, even when compared to the first approach that require more resources, such as parallel corpora or bilingual dictionaries. The authors’ future work will include the identification of other language-independent methods that could provide feature selection to improve clustering.

This work provides improvements to multilingual information retrieval. Indeed, language-independent document clustering facilitates the identification of all documents about the same subject.

Charton and Torres-Moreno's "Modélisation automatique de connecteurs logiques par analyse statistique du contexte" is more technical and deals with the identification of certain types of synonym, viz logical connectors (*donc, en conséquence, par conséquent*, etc.) that are interchangeable to some degree.

The method presented is based on contexts shared by different logical connectors. When sufficiently many contexts are repeated for two logical connectives, they are deemed synonyms. This echoes similar work on synonym identification, but addresses content words and not stop words, such as logical connectors (adverbs, or adverbial or prepositional phrases).

The method is presented and evaluated on various texts (Senate debates, literary texts). The results evaluated were composed of contexts in which an equivalent of a logical connector was automatically substituted for another. An expert evaluated the accuracy of the result, and substitutions were evaluated as correct between 79% and 88% of cases, depending on the type of text.

The first application of this work to information retrieval would be to dissect the particular sentence structure and help the identification of target phrases. The authors' perspective of applying their method to synonym substitution suggests further applications to information retrieval.

The article "A sentiment-based digital library of movie review documents using Fedora" by Na and colleagues presents a digital library of film reviews, which allows searches based on the opinions expressed in the reviews (positive, negative, or neutral) on various aspects of films. Not only can a movie title be found, but also films for which the director's work was judged favourably, or movies considered to be poor. Access is thus based both on traditional metadata fields and "emotional" aspects of the documents, obtained by automatic content analysis.

First, to allow classification of analyses of sentiments about different perspectives on films (director, cast, or film as a whole), phrases in film reviews were labelled according to these aspects. Labelling was performed using information-extraction techniques, including identification of named entities, identification of co-reference between two expressions (e.g., the fact that Carrey and Jim Carrey denote the same person), and pronoun resolution. Then the classification of the reviews based on these aspects was performed by a supervised machine-learning algorithm (a support vector machine). The authors present an assessment of the accuracy of

this classification system ranging between 69.1% and 90.48%, depending on the aspect considered.

The article also describes the digital library of movie reviews that was developed. It allows browsing and searching based on the sentiments towards the director, the cast, and the entire film. The library has a web interface and a database (stored in the Fedora application) that contains the results of the sentiment analysis process and the classification system described in the first part of the article. The search interface contains fields corresponding to the selected metadata (title of film, cast, director). Entries corresponding to each review are each stored under a different tab, allowing the user to read only specific parts. In a query, the user can choose to read only positive, negative, or neutral opinions, or all reviews. Interface elements (a hand with a thumb up or down) synthesize the judgements and facilitate navigation through the results.

This work presents a complete system, from document collection to annotation, and the creation of the search interface and navigation tool. It clearly illustrates how information-extraction techniques support information retrieval, which can take into account the emotional aspects of the documents.

The field of text mining is an active research area, and its applications are numerous (information discovery, sentiment analysis, topic spotting, etc.). This special issue focuses, however, on the latest research developments that integrate text-mining techniques to assist information retrieval. In addition to presenting recent work in this area, we believe that this volume will help readers to understand the relevance of combining text-mining and information-retrieval techniques.

References

- Feldman, R., and J. Sanger. 2006. *The text mining handbook: Advanced approaches in analysing unstructured data*. Cambridge: Cambridge University Press.
- Ibekwe-SanJuan, F. 2007. *Fouille de textes : méthodes, outils et applications*. Paris : Hermès.
- Srivastava, A., and M. Sahami, eds. 2009. *Text mining: Classification, clustering, and applications*. Boca Raton: CRC Press.
- Weiss, S.M., N. Indurkha, T. Zhang, and F.J. Damerau. 2005. *Text mining: Predictive methods for analyzing unstructured information*. Berlin: Springer-Verlag.