INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 3 ISSUE 4 (2017) PAGES 2549 - 2553
RECEIVED : 05.07.2017 PUBLISHED : 03.08.2017
Aug 03, 2017

IMPROVING THE TEXT MINING PROCESS WITH EFFECTIVE PATTERN CATEGORIZATION MODEL

K.Jayanthi[1], C.Karthi[2] ,D.B.Shanmugam[3]

[1]M.Phil, Research Scholar, Dr.M.G.R.Chockalingam Arts College, Arni.

[2]Assistant Professor, Department of MCA, Sri Balaji Chockalingam Engineering College , Arni.

[3]Associate Professor, Department of MCA, Sri Balaji Chockalingam Engineering College , Arni.

ABSTRACT

The enormous amount of information stored in unstructured texts cannot simply be used for further processing by computers, which typically handle text as simple sequences of character strings. Therefore, specific pre-processing methods and algorithms are required in order to extract useful patterns. Text mining refers generally to the process of extracting interesting information and knowledge from unstructured text. In this thesis, we discuss text mining as a young and interdisciplinary field in the intersection of the related areas information retrieval, machine learning, statistics, computational linguistics and especially data mining.

Text data typically consists of strings of characters, which are transformed into a representation suitable for learning. It is observed from previous research that words work well as features for many text categorization tasks. In the feature space representation, the sequences of characters of text documents are represented as sequence of words. Feature selection involves tokenizing the text, indexing and feature space reduction. Text can be tokenized using term frequency (TF), inverse document frequency (IDF), term frequency inverse document frequency (TFIDF) or using binary representation. Using these representations the global feature space is determined from entire training document collection.

Many data mining techniques have been proposed for mining useful patterns in text documents. However, how to effectively use and update discovered patterns is still an open research issue, especially in the domain of text mining. Over the years, people have often held the hypothesis that pattern (or phrase)-based approaches should perform better than the term-based ones, but many experiments do not support this hypothesis. This thesis presents an innovative and effective pattern discovery technique which includes the processes of pattern deploying and pattern evolving, to improve the effectiveness of using and updating discovered patterns for finding relevant and interesting information. Substantial experiments on RCV1 data collection and TREC topics demonstrate that the proposed solution achieves encouraging performance.

1. INTRODUCTION

DUE to the rapid growth of digital data made available in recent years, knowledge discovery and data mining have attracted a great deal of attention with an imminent need for turning such data into useful information and knowledge. Many applications, such as market analysis and business management, can benefit by the use of the information and knowledge extracted from a large amount of data. Knowledge discovery can be viewed as the process of nontrivial extraction of information

INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 3 ISSUE 4 (2017) PAGES 2549 - 2553
RECEIVED : 05.07.2017 PUBLISHED : 03.08.2017
Aug 03, 2017

from large databases, information that is implicitly presented in the data, previously unknown and potentially useful for users. Data mining is therefore an essential step in the process of knowledge discovery in databases.

In the past decade, a significant number of data mining techniques have been presented in order to perform different knowledge tasks. These techniques include association rule mining, frequent item set mining, sequential pattern mining, maximum pattern mining, and closed pattern mining. Most of them are proposed for the purpose of developing efficient mining algorithms to find particular patterns within a reasonable and acceptable time frame. With a large number of patterns generated by using data mining approaches, how to effectively use and update these patterns is still an open research issue. We focus on the development of a knowledge discovery model to effectively use and update the discovered patterns and apply it to the field of text mining. Text mining is the discovery of interesting knowledge in text documents. It is a challenging issue to find accurate knowledge (or features) in text documents to help users to find what they want. In the beginning, Information Retrieval (IR) provided many term-based methods to solve this challenge, such as Rocchio and probabilistic models, rough set models , BM25 and support vector machine (SVM)based filtering models. The advantages of termbased methods include efficient computational performance as well as mature theories for term weighting, which have emerged over the last couple of decades from the IR and machine learning communities. However, term based methods suffer from the problems of polysemy and synonymy, where polysemy means a word has multiple meanings, and synonymy is multiple words having the same meaning. The semantic meaning of many discovered terms is uncertain for answering what users want.

2. LITERATURE REVIEW

Text Mining is the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources. A key element is the linking together of the extracted information together to form new facts or new hypotheses to be explored further by more conventional means of experimentation. Text mining is different from what are familiar with in web search. In search, the user is typically looking for something that is already known and has been written by someone else. The problem is pushing aside all the material that currently is not relevant to your needs in order to find the relevant information. In text mining, the goal is to discover unknown information, something that no one yet knows and so could not have yet written down.

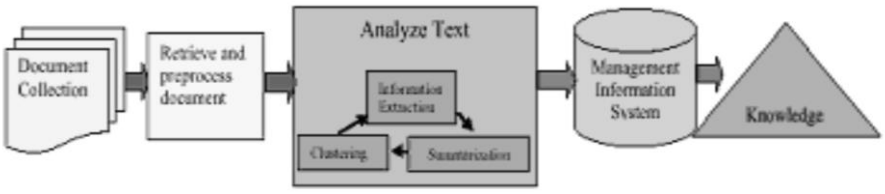

Fig.1. Example of text mining

INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 3 ISSUE 4 (2017) PAGES 2549 - 2553
RECEIVED : 05.07.2017 PUBLISHED : 03.08.2017
Aug 03, 2017

Although the differences in human and computer languages are expansive, there have been technological advances which have begun to close the gap. The field of natural language processing has produced technologies that teach computers natural language so that they may analyze, understand, and even generate text. A topic tracking system works by keeping user profiles and, based on the documents the user views, predicts other documents of interest to the user. Yahoo offers a free topic tracking tool (www.alerts.yahoo.com) that allows users to choose keywords and notifies them when news relating to those topics becomes available. Topic tracking technology does have limitations, however. For example, if a user sets up an alert for "text mining", s/he will receive several news stories on mining for minerals, and very few that are actually on textmining. Some of the better text mining tools let users select particular categories of interest or the software automatically can even infer the user's interests based on his/her reading history and click-through information.

3. TEXT MINING APPLICATIONS

The sectors analyzed are characterized by a fair variety in the applications being experimented. however, it is possible to identify some sectorial specifications in the use of TM, linked to the type of production and the objectives of the knowledge management leading them to use TM. The publishing sector, for example, is marked by prevalence of Extraction Transformation Loading applications for the cataloguing, producing and the optimization of the information retrieval.

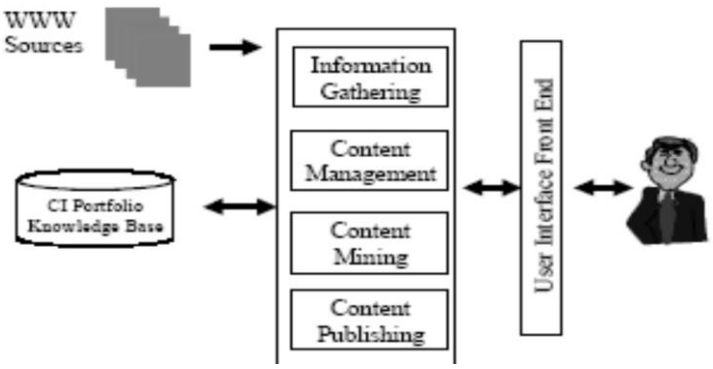

Fig.2. FOCI System Architecture

In the banking and insurance sectors, on the other hand, CRM applications are prevalent and aimed at improving the management of customer communication, by automatic systems of message re-routing and with applications supporting the search engines asking questions in natural language. In the medical and pharmaceutical sectors, applications of Competitive Intelligence and Technology Watch are widespread for the analysis, classification and extraction of information from articles, scientific abstracts and patents. A sector in which several types of applications are widely used is that of the telecommunications and service companies: the most important objectives of these industries are that all applications find an answer, from market analysis to human resources management, from spelling correction to customer opinion survey.

4. RESULT ANALYSIS

Many types of text representations have been proposed in the past. A well known one is the bag of words that uses keywords (terms) as elements in the vector of the feature space. The tf*idf

INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 3 ISSUE 4 (2017) PAGES 2549 - 2553
RECEIVED : 05.07.2017 PUBLISHED : 03.08.2017
Aug 03, 2017

weighting scheme is used for text representation in Rocchio classifiers. In addition to TFIDF, the global IDF and entropy weighting scheme is proposed in and improves performance by an average of 30 percent. Various weighting schemes for the bag of words representation approach were given in . The problem of the bag of words approach is how to select a limited number of features among an enormous set of words or terms in order to increase the system's efficiency and avoid overfitting.

In data mining techniques have been used for text analysis by extracting cooccurring terms as descriptive phrases from document collections. However, the effectiveness of the text mining systems using phrases as text representation showed no significant improvement. The likely reason was that a phrase-based method had "lower consistency of assignment and lower document frequency for terms" as mentioned. Term-based ontology mining methods also provided some thoughts for text representations. For example, hierarchical clustering was used to determine synonymy and hyponymy relations between keywords. Also, the pattern evolution technique was introduced in order to improve the performance of term-based ontology mining. Pattern mining has been extensively studied in data mining communities for many years.

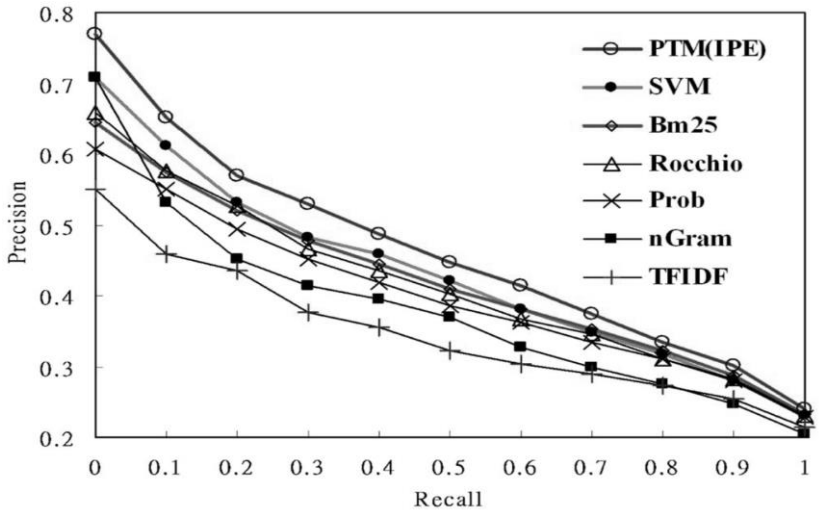

Fig.3. Output Analysis

These results strongly support Hypothesis H1. The promising results can be explained in that the use of the deploying method is promising (Hypothesis H2 is also supported) for solving the misinterpretation problem because it can combine well with the advantages of terms and discovered patterns or concepts. Moreover, the inner pattern deploying strategy provides an effective evaluation for reducing the side effects of noisy patterns because the estimation of term weights in the term space is based on not only terms' statistical properties but also patterns' associations in the corresponding pattern taxonomies.

CONCLUSION

Many data mining techniques have been proposed in the last decade. These techniques include association rule mining, frequent itemset mining, sequential pattern mining, maximum pattern mining, and closed pattern mining. However, using these discovered knowledge (or patterns) in the

field of text mining is difficult and ineffective. The reason is that some useful long patterns with high specificity lack in support (i.e., the low-frequency problem). We argue that not all frequent short patterns are useful. Hence, misinterpretations of patterns derived from data mining techniques lead to the ineffective performance. In this research work, an effective pattern discovery technique has been proposed to overcome the low-frequency and misinterpretation problems for text mining. The proposed technique uses two processes, pattern deploying and pattern evolving, to refine the discovered patterns in text documents. The experimental results show that the proposed model outperforms not only other pure data mining-based methods and the concept based model, but also term-based state-of-the-art models, such as BM25 and SVM-based models.

REFERENCES

[1] K. Aas and L. Eikvil, "Text Categorisation: A Survey," Technical Report Raport NR 941, Norwegian Computing Center, 1999.

[2] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules in Large Databases," Proc. 20th Int'l Conf. Very Large Data Bases (VLDB '94), pp. 478-499, 1994.

[3] H. Ahonen, O. Heinonen, M. Klemettinen, and A.I. Verkamo, "Applying Data Mining Techniques for Descriptive Phrase Extraction in Digital Document Collections," Proc. IEEE Int'l Forum on Research and Technology Advances in Digital Libraries (ADL '98), pp. 2-11, 1998.

[4] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval. Addison Wesley, 1999.

[5] N. Cancedda, N. Cesa-Bianchi, A. Conconi, and C. Gentile, "Kernel Methods for Document Filtering," TREC, trec.nist.gov/ pubs/trec11/papers/kermit.ps.gz, 2002.

[6] N. Cancedda, E. Gaussier, C. Goutte, and J.-M. Renders, "Word- Sequence Kernels," J. Machine Learning Research, vol. 3, pp. 1059- 1082, 2003.

[7] M.F. Caropreso, S. Matwin, and F. Sebastiani, "Statistical Phrases in Automated Text Categorization," Technical Report IEI-B4-07- 2000, Instituto di Elaborazione dell'Informazione, 2000.