GNITED MINDS
Journals

International Journal of Information Technology and Management

Vol. VIII, Issue No. XI, February-2015, ISSN 2249-4510

EFFICIENT PREPROCESSING AND PATTERN IDENTIFICATION APPROACH FOR TEXT MINING

www.ignited.in

International Journal of Information Technology and Management
Vol. VIII, Issue No. XI, February-2015, ISSN 2249-4510

Efficient Preprocessing and Pattern Identification Approach for Text Mining

Amreen Ahmad

Assistant Professor - BBDIT

Abstract – Many data processing techniques are proposed for mining helpful patterns in text documents. However, how to effectively use and update those discovered patterns remains an open analysis issue, particularly within the domain of text mining. Pattern mining is an important research issue in data mining with few kinds of applications. In text documents, a significant number of data mining techniques have been proposed for mining useful patterns. But there are some questions; how to effectively use and update discovered patterns is still an open research issue in the area of text mining. Many text mining methods have been proposed for mining useful pattern in text documents.

---------------------------♦-----------------------------

INTRODUCTION

Text mining is a new area of computer science that fosters strong connections with natural language processing, data mining, machine learning, information retrieval and knowledge management. Text mining seeks to extract useful information from unstructured textual data through the identification and exploration of interesting patterns [2]. It mainly focuses to approximate and identify different entities such as terms, phrases and pattern. Then the system assigns the frequency to each word, all the weight of the document is used for pattern clustering. Pattern clustering is one of the favorable methods for feature extraction in text classification. In this paper we propose a fuzzy estimated and similarity - based self-generating algorithm for text classification.

Due to the rapid increase of digital data made available recently, knowledge discovery and data mining [1] have attracted a large amount of attention which includes an imminent need for turning such data into useful suggestions and knowledge. Many applications, for instance market analysis and business, may benefit by way of the information and knowledge extracted from a considerable amount of data. Knowledge discovery can be viewed as the method of nontrivial extraction of real info from large databases, information that's implicitly presented among the data, previously unknown and potentially ideal for users. Data mining is therefore of vital help in the method of knowledge discovery in databases. A decade before, a major wide range of data mining techniques have been presented in an effort 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.

REVIEW OF LITERATURE:

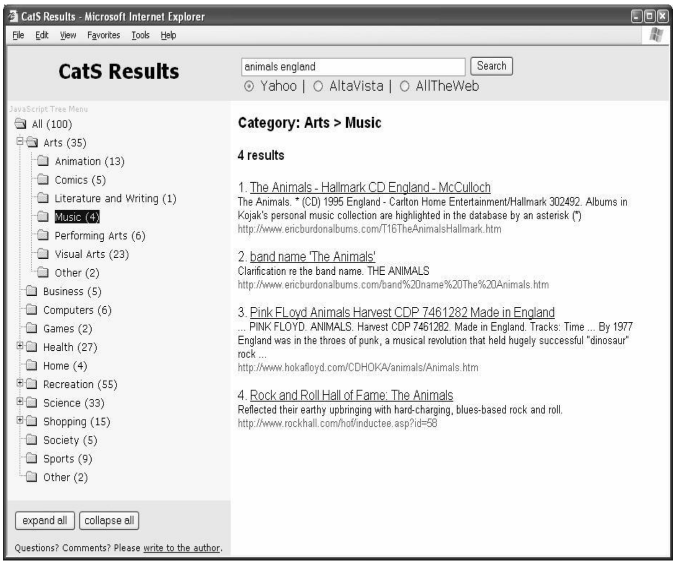
The challenging issue is how to effectively content with the big quantity of discovered patterns. Regarding the challenging issue, closed sequential designs could have been utilized for text mining in [4], which proposed that the idea of closed patterns in text mining was, useful in addition to the possibility for improving the appearance of the performance of message mining. Pattern taxonomy model was also developed in [1] and [2] to further improve the effectiveness by effectively using closed patterns in text mining. Additionally, a two-stage model that used both term-based methods and pattern based methods made its entrance in [3] to significantly improve the performance of real information filtering. Natural language processing (NLP) serves as a modern computational technology that in fact can assist individuals to understand the meaning of message documents. For a very long time, NLP was struggling for grappling with uncertainties in human languages. Recently, a new concept-based model [2], [1] was presented to bridge the gap between NLP and text mining, which analyzed terms on the sentence and document levels. Pattern based techniques was introduced in [2] to significantly improve the performance of information filtering.

BAG-OF-WORDS DOCUMENT REPRESENTATION:

Let W be the dictionary – the set of all terms (words) that occur at least once in a collection of documents D. The bag-of-words representation of document dn is a vector of weights (w1n,. . . ,w|W|n). In the simplest case, the weights win ∈ {0, 1} and denote the presence or absence of a particular term in a document. More commonly, win represent the frequency of the ith term in the nth document, resulting in the term frequency representation. Normalization can be employed to scale the term frequencies to values between 0 and 1, accounting for differences in the lengths of documents. Besides words, n-grams may also be used as terms. However, two different notions have been referred to as “n-grams” in the literature. The first are phrases as sequences of n words, while the other notion are n-grams as sequences of characters. N-grams as phrases are usually used to enrich the BOW representation rather than on their own. N-grams as sequences of characters are used instead of words [6].

ENHANCING WEB SEARCH:

One way to enhance users’ efficiency and experience of Web search is by means of meta-search engines. Traditionally, meta-search engines were conceived to address different issues concerning general-purpose search engines, including Web coverage, search result relevance, and their presentation to the user. A common approach to alternative presentation of results is by sorting them into (a hierarchy of) clusters which may be displayed to the user in a variety of ways, e.g. as a separate expandable tree (vivisimo.com) or arcs which connect Web pages within graphically rendered “maps” (kartoo.com). However, topics generated by clustering may not prove satisfactory for every query, and the “silver bullet” method has not yet been found. An example of a meta-search engine that sorts search results into a hierarchy of topics using text categorization techniques is CatS [7] (stribog.im.ns.ac.yu/cats). Figure 1 shows the subset of the 100 results for query ‘animals england’ sorted into category Arts → Music, helping separate pages about animals living in England from pages concerning the English music scene. The categories employed by CatS were extracted from the dmoz Open Directory.


CatS Results - Microsoft Internet Explorer
File Edit View Favorites Tools Help
CatS Results
Search
Yahoo | AltaVista | AllTheWeb
JavaScript Tree Menu
expand all
collapse all

animals england

● Yahoo | ○ AltaVista | ○ AllTheWeb

All (100)
Arts (35)
Animation (13)
Comics (5)
Literature and Writing (1)
Music (4)
Performing Arts (6)
Visual Arts (23)
Other (2)
Business (5)
Computers (6)
Games (2)
Health (27)
Home (4)
Recreation (55)
Science (33)
Shopping (15)
Society (5)
Sports (9)
Other (2)

Category: Arts > Music

4 results

1. The Animals - Hallmark CD England - McCulloch
The Animals. * (CD) 1995 England - Carlton Home Entertainment/Hallmark 302492. Albums in Kojak's personal music collection are highlighted in the database by an asterisk (*)
http://www.ericburdonalbums.com/T16TheAnimalsHallmark.htm

2. band name 'The Animals'
Clarification re the band name. THE ANIMALS
http://www.ericburdonalbums.com/band%20name%20The%20Animals.htm

3. Pink FLoyd Animals Harvest CDP 7461282 Made in England
... PINK FLOYD. ANIMALS. Harvest CDP 7461282. Made in England. Tracks: Time ... By 1977 England was in the throes of punk, a musical revolution that held hugely successful "dinosaur" rock ...
http://www.hokafloyd.com/CDHOKA/animals/Animals.htm

4. Rock and Roll Hall of Fame: The Animals
Reflected their earthy upbringing with hard-charging, blues-based rock and roll.
http://www.rockhall.com/hof/inductee.asp?id=58

Questions? Comments? Please write to the author.

Figure 1: Results for query ‘animals england’ classified into Arts ! Music (Source from [8,6])

CONCLUSION:

Many data mining techniques have been proposed in the last decade. These techniques include sequential pattern mining, maximum pattern mining, association rule mining, frequent item set mining,, and closed pattern mining[4].

The quality of extracted features is the key issue to text mining due to the large number of terms, phrases, and noise. Most existing text mining methods are based on term-based approaches which extract terms from a training set for describing relevant information. However, the quality of the extracted terms in text documents may be not high because of lot of noise in text. 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).

REFERENCES:

[1] 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.

[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] Hybrid Approach to Improve Pattern Discovery in Text mining Charushila Kadu, International Journal of Advanced Research in Computer and Communication Engineering

[4] Effective Pattern Discovery for Text Mining Ning Zhong, IEEE Transactions on knowledge and data engineering, VOL. 24, NO. 1,

[5] Pattan Kalesha, M. Babu Rao,Ch. Kavitha, Efficient Preprocessing and Patterns Identification Approach for Text Mining, International Journal of Computer Trends and Technology (IJCTT) – volume 6 number 2– Dec 2013

[6] Milos Radovanovic Mirjana Ivanovic, Text mining: approaches and applications, Novi Sad J. Math. Vol. 38, No. 3, 2008, 227-234

[7] Radovanović M., Ivanović, M., CatS: A classification-powered meta-search engine. In: Last, M., et al., editors, Advances in Web

Intelligence and Data Mining, pages 191–200, Springer-Verlag, 2006.

[8] http://www.emis.de/journals/NSJOM/Papers/38_3/NSJOM_38_3_227_234.pdf