International Journal Of Engineering Research And Development
e- ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com
Volume 11, Issue 01 (January 2015), PP. 70-72

Classification and Distribution of E-News Paper

Ganesh J Palve
DepartmentofComputerScience&Engineering
G.H.RIEMJalgaon

ABSTRACT
In many real-world scenarios, the ability to automatically classify documents into a fixed set of categories is highly desirable. Common scenarios include classifying a large amount of unclassified archival documents such as newspaper articles, legal records and academic papers. For example, newspaper articles can be classified as 'features', 'sports' or 'news'. Other scenarios involve classifying of documents as they are created. Examples include classifying movie review articles into 'positive' or 'negative' reviews or classifying only blog entries using a fixed set of labels. Natural language processing offers powerful techniques for automatically classifying documents. These techniques are predicated on the hypothesis that documents in diffrent categories distinguish themselves by features of the natural language contained in each document. Salient features for document classification may include word structure, word frequency, and natural language structure in each document.

PROBLEM STATEMENT:

Ourproposed projectis automatically classifying newspaper articles from the Different newspaper. The Newspaper has archives of a large number of articles which require classification into specific sections (News, Opinion, Sports, etc). Our project is aimed finding and building techniques which can be used to perform automatic articles classification. At our disposal is a large archive of already classified documents so we are able to make use of supervised classification techniques. We randomly divide this archive of classified documents into groups like training and testing for our classification systems. This project experiments with different NLP feature sets as well as different statistical techniques using these feature sets and compares the performance in each case. Specifically, our proposed project involves experimenting with feature sets for Naive Bayes Classification, Maximum Entropy Classification, and examining sentence structure differences in different categories using probabilistic grammar parsers.

I. INTRODUCTION:

Classification of data is used to find the category to which information or data belongs. In the present situation, due to technology and information, NEWS is easily accessible through online sources. Now a days different Data is accessed by users through many sources like information media, computer media and many more. Normallythese dataare not available in organized form and but it can be converted to aorganized form. To convert these unorganized data into Particular structured form Data Pre-Processing techniques are used(like punctuation removal, stop word removal, stemming, lemmatizing the words, removing special characters) and further it is classified. In this paper various data classification models like Naïve Bayes,SVM(Support Vector Machine), and Logistic regression are compared to identify best module which gives accurate results of classification of NEWS.

This paper is divided intoDifferent sections

(I. Introduction,
(II. An overview of TDOCS Toolkit,
(III. Thesaurus Technology,
(IV. Application
(V. Features
(VI. Conclusions.

Section II is a general introduction to Thesaurus based Documents Classification System "TDOCS" Toolkit.
Section III the thesaurus technology, discusses the role of thesaurus in information retrieval and describes how the E-Newspaper Classification and Distribution Based on User Profiles and Thesaurus has been created and used to support the classification and user profiling functions.
Section IV Gives the application
In Section V Feature Scope.
The last section presents the conclusions.

II. AN OVERVIEW OF TDOCS

Electronic thesauri have been identified as strategic instruments for indexing electronic documents. However, one of the main problems of using electronic thesauri remains the creation and maintenance. In this section, we provide some basic information towards a solution for this maintenance problem through the TDOCS-Toolkit and ThesWB. The concept of the TDOCS document management system is shown in Fig. 1. Electronic documents are imported into the system via the indexing process.

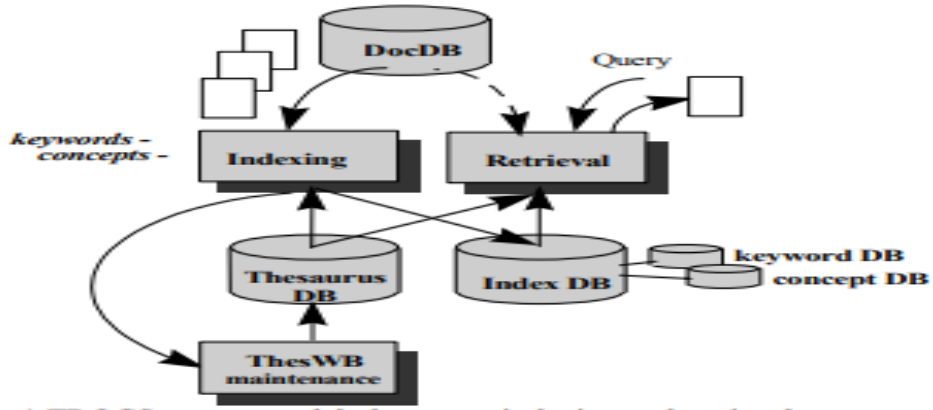

Fig. 1:TDOCS system model: document indexing and retrieval.

This process generates automatically various index terms such as keywords, concepts and relations, which are assigned to documents as a function of the document content. Thesaurus DB Index DB Indexing Retrieval Query DocDB keywords - concepts - keyword DB concept DB ThesWB maintenance Fig. 1: TDOCS system model: document indexing and retrieval. The indexing and retrieval processes being used in TDOCS are based on one part on the hierarchical structure of the thesaurus. The thesaurus hierarchy is used in order to create associations between documents and concepts. The use of the thesaurus therefore implies that the associations can also be expressed by terms not explicitly present in the analyzed text. Such terms, we call concepts; refer to broader terms of keywords that are found in the document. The indexing process usually identifies in documents not only the keywords, but also high-frequency terms, which are not present in the thesaurus. In spite of their relevancy as index terms, those terms are not indicated as keywords because of their absence in the thesaurus. To incorporate those terms in 1 TDOCS is bsed on IKEM - method. the thesaurus, an additional iteration is needed, which is an update of the thesaurus with those high-frequency terms using ThesWB Tool. This iteration needs to be followed by reindexing of documents in order to maintain the consistency of the index and thesaurus databases.

III. THESAURUS TECHNOLOGY.

Thesauri and Information Retrieval Thesaurus technology has been used to assist in the retrieval of information for more four decades. Although free text retrieval has become quite popular during the last decade, the use of thesauri as a component of information retrieval system is challenging. Thesauri are tools aimed of improving the effectiveness of information retrieval system. First, they can assist term selection through the semantic roadmap they provide and second, they can be used for automatic query expansion. Thesaurus technology undoubtedly indicates an innovative approach in document management, especially in the field of document indexing and retrieval. Every day millions of people are searching every second for information stored in documents somewhere locally, in an intranet or on the Internet, in order to find the specific information they want to find that at specific time. In general, indexing documents based on a thesaurus means looking for a match between thesaurus terms and terms occurring in the document to be indexed. Each time there is a match the thesaurus term is assigned as a key term to the document. Another characteristic feature of thesaurus based indexing is using concept for indexing allows document to be ranked as a relevant to a query, even if the query term itself does not occur in the text, but only a related term which denotes the same concept. Applied to document retrieval this means that thesaurus based indexing allows documents to be retrieved even if one or more of some given words in a search string do not match to any word or combination of words in the documents.

IV. APPLICATION

- Today, people publish millions of messages per day on Twitter, which is the most popular micro blogging service on the Web. Recent research shows that the majority of Twitter messages (tweets) are related to news and that the trending topics propagate quickly through the Twitter social network which allows for applications such as early warning systems. So far, most research initiatives focused on the analysis of structural properties of the Twitter network. The proposed work can be extended to how individual Twitter activities can be exploited to infer personal interests and generate semantic user profiles that can be re-used also by other applications than Twitter.
- In contrast to other Social Web services like Last.fm, which allows for the deduction of users' musical tastes, or Flicker, which primarily provides information to infer users' interests in locations or events, tweets are not restricted to a certain domain. Instead, Twitter users can discuss about any topic they are interested in or concerned with which makes it worthwhile to explore for user modeling. Furthermore, the real-time nature of information that people publish on Twitter poses new challenges and possibilities for user modeling.
- In Media, the goal of a documentation department is to help journalists to find information in archived news in order to reuse it in a new article. To this end, these departments have to tag news every day, and the typical way to do that is by using a thesaurus: a set of items (word or phrases) used to classify things.

V. FUTURESCOPE

Finding newspaper articles coverage of events can be complicated. Different indexing and searching resources are needed depending on the date and the geographic location of the event and perspective desired. Public indexing of many newspapers is relatively recent phenomena, complicating matters further. More effort is needed to focus on these issues.

To summarize, automatic E-Newspaper classification is an important problem nowadays.

Here we propose an approach base on tf-idf and TDOCS indexing platform is not only an indexing tool, but also a classification tool. to classify and distribute the E-Newspaper articles.

VI. CONCLUSION:

The International Press Thesaurus is useful and effective for indexing and retrieval of electronic newspaper articles. Concepts hierarchies in International Press Thesaurus was used to capture user profile and classify E-Newspaper articles content. Our experiment proves that TDOCS indexing platform is not only an indexing tool, but also a classification tool. ThesWB Tool is a useful tool to update or create a thesaurus.

REFERENCES

[1]. Guarino, N., Masolo, C., Vetere, G.: OntoSeek: Content-Based Access to the Web. IEEE Intelligent Systems 14(3), 70{80 (1999).

[2]. Sergio Cleger-Tamayo , Juan M. Fernández-Luna , Juan F. Huete "Knowledge-Base System- 'Top-N news recommendations in digital newspapers' ", the research programme ConsoliderIngenio 2010: MIPRCV (CSD2007-00018) and the Junta de Andalucía, Page 180- 189, 2011

[3]. Ahn, J. ,Brusilovsky, P., Grady, J., He, D., Syn, S.Y.: Open User Profiles for Adaptive News Systems: Help or Harm? In: 16th International Conference on World Wide Web (WWW 2007). pp. 11-20. ACM (2007)

[4]. IJntema W., Goossen F., Frasincar F., Hogenboom F.: Ontology-Based News Recommendation. EDBT 2010, March 22–26, 2010, Lausanne, Switzerland. Copyright 2010 ACM 978-1-60558-945-9/10/0003

[5]. Billsus, D., Pazzani, M.J.: User Modeling for Adaptive News Access. User Modeling and User-Adapted Interaction 10(2), 147-180 (2000)

[6]. Guarino, N., Masolo, C., Vetere, G.: OntoSeek: Content-Based Access to the Web. IEEE Intelligent Systems 14(3), 70-80 (1999)

[7]. Salton, G., Buckley, C.: Term-Weighting Approaches in Automatic Text Retrieval. Information Processing and Management 24(5), 513-523 (1988)

[8]. Baziz, M., Boughanem, M., Traboulsi, S.: A Concept-Based Approach for Indexing Documents in IR. IRIT, Campus universitaireToulouseIII 118 rte de Narbonne, F-31062 Toulouse Cedex 4, France {baziz, boughane, traboul@irit.fr

[9]. Cantador, I., Bellogn, A., Castells, P.: News@hand: A Semantic Web Approach to Recommending News. EscuelaPolitécnica Superior, Universidad Autónoma de Madrid Campus de Cantoblanco, 28049 Madrid, Spain

[10]. Billsus, D., Pazzani, M.J.: 'A Personal News Agent that Talks, Learns and Explains, Department of Information and Computer Science University of California, Irvine Irvine, CA 92697 +1 (949) 824-3491.

[11]. Houda El Bouhissi, Mimoun Malki and Djamila Berramdane, "Applying Semantic Web Services", International Journal of Computer Engineering & Technology (IJCET), Volume 4, Issue 2, 2013, pp. 108 - 113, ISSN Print: 0976 – 6367, ISSN Online: 0976 – 6375.

[12]. S Prerna, Sanjay Singh, Rajesh Singh and Monika Jena, "Interactive News Feed Extraction System", International Journal of Computer Engineering & Technology (IJCET), Volume 4, Issue 2, 2013, pp. 10 - 16, ISSN Print: 0976 – 6367, ISSN Online: 0976 – 6375.