A survey on Information Retrieval System for Online Newspapers

Dr. M Hanumanthappa[1], Deepa T. Nagalavi[2]

1) *Associate Professor, Dept. of Computer Science and Applications, Jnana Bharathi Campus, Bangalore University, Bangalore -560 056, INDIA*
2) *Research Scholar, Department of Computer Science and Applications, Bangalore University, Bangalore-56*

E-mail:- hanu6572@hotmail.com[1], deepanagalavi@gmail.com[2]

Abstract

"Anywhere, anytime" is the buzz word of the present era. The World Wide Web has been instrumental in providing rich source of information anywhere, any time. Efficiently searching and effectively retrieving this information is a challenge because it not only involves text data but also data in multimedia format. Online news is commonly read by many users. Categorizing and compiling news from various newspapers, ranking the retrieved news based on the content and presenting the information is an important task. In this paper we have done a survey of categorizing the news with various Data Mining techniques. Ranking and compiling the news items based on a query is presented

.

Keywords: Information Retrieval, Data mining, clustering.

1. Introduction

The events that are happening around us reaches the common man in digitized form through newspapers on a daily basis. In earlier days newspapers were printed on relatively inexpensive, low-grade paper. In this present information era, the news updates are presented online through the web as E-Newspapers that resemble exactly the print papers. As there are multiple news agencies, the user has a variety of newspapers at click of a button to navigate and read. This is a time consuming task. Thus the goal is to retrieve the related information efficiently from different online newspapers for a particular topic irrespective of the format, integrate it and present it to the users in a suitable form. Information Retrieval supported by Data Mining techniques can be used to achieve the goal.

Information Retrieval (IR) is the process by which a collection of data is represented, stored, and searched for the purpose of knowledge discovery as a response

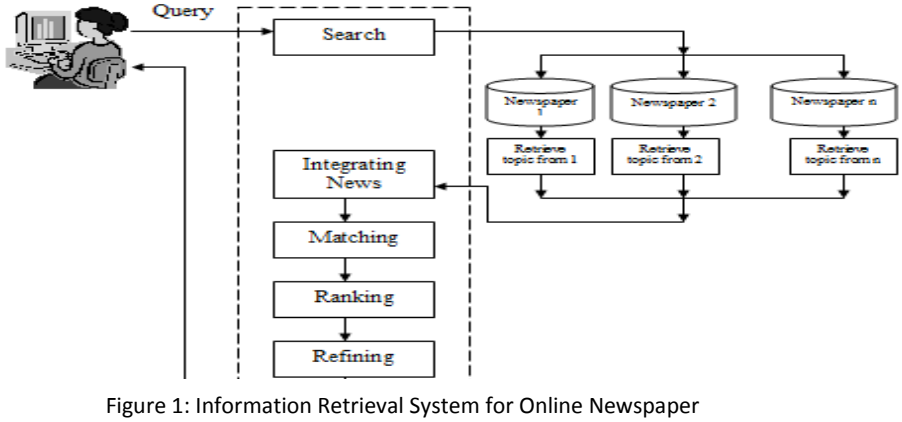

Figure 1: Information Retrieval System for Online Newspaper

to a user request (query). This process involves various stages starting with representing data and ending with returning relevant information to the user. Intermediate stages include, filtering, searching, matching and ranking operations. Most IR systems compute a numeric score on how well each object in the database matches the query, and rank the objects according to this value. The top ranking objects are then shown to the user. The process may then be iterated if the user wishes to refine the query. The process flow of IR System for online Newspaper is shown in figure 1. IR is concerned with the organization and retrieval of information from a large number of text-based documents [9].

A general IR system usually carries out some processing on the user request to derive a form of the request that it can match directly against the document collection using some form of matching algorithm. The processed request, which may take many forms, is known as the query. Query formats commonly employed in the IR world include the natural language query, where the request is not processed much at all, and the bag of words format, where function words, punctuation and phrases like “on the subject of" are removed from the request, suffixes stripped, and a selection of what are known as keywords extracted to form the query.

The query is matched against the document collection using a matching algorithm which calculates a score for each document in the collection reacting its perceived similarity to the query. Similarity scores may be based simply on the frequency of individual query terms (words or phrases), or may exploit term weights (scores per term) calculated using frequency data. The Vector Space Method and the Probabilistic Model of Information Retrieval (which is the model employed by the IR system used in the experiments discussed in subsequent chapters) provide well-founded ways of doing this. Generally, a list of the N most closely matching documents is returned to the user. This list of returned documents is often called the retrieved document list. The aim is to retrieve as many relevant documents as possible in this list, while avoiding the retrieval of irrelevant ones.

IR systems are generally evaluated using two metrics, precision and recall. Precision is defined as the proportion of retrieved documents which are actually relevant to the query derived from the user request.

Precision = Not Relevant / Total Retrieved

Recall is the proportion of documents known to be relevant to the query in the entire collection that have been retrieved in the retrieved document list for that query.

Recall = Not Relevant Retrieved / Total Known Relevant

Many IR system is based on Data Mining techniques. Data Mining is a term which deals with large amount of data. It is a process of discovering interesting knowledge from large amount of data stored in different information repositories namely databases, and data warehouse [3]. Data Mining Techniques include Clustering, Classification, Association, Trend Analysis among these clustering and classification are widely used to categorize retrieved news. Clustering has been proven to be a useful technique for IR. A web search engine often returns enormous amount of information in response to a broad query, making it difficult for users to browse or to identify relevant information. With the help of clustering methods we can automatically group the retrieved information into a list of meaningful categories [5]. Online news can be categorized into various categories sports, travel, technology, business, entertainment, and many more. Other than clustering, another techniques known as text classification can be used to classify news into different categories. Online news papers provide news under topical categories like national, international, politics, finance, sports, entertainment etc. News article on topical issue are helpful for company managers and other decision makers. Classification start with training set of documents that are labeled with class label. Text classification is classified into two categories (i) single label (ii) multi label. A single label belongs to only one class where as a multi label belongs to more than one class [5].

This paper presents a survey on Information Retrieval System for online Newspapers.

2. Literature Review

Retrieving news directly from newspapers can result time expensive for users because many news papers are specialized in one kind of news. We can find search tools like google, yahoo, altavista, etc to retrieve news from different newspapers but some of them don’t provide such an efficient result, because they retrieve too many documents of which some are relevant to the users query and most relevant documents are not in order. IR provides a mechanism to rank the retrieved objects based on query [10].

In the literature some IR tasks have been proposed, Such as News Miner[1], News retrieval through a multi agent system[4]. Alberto Sillitti Marco Scotto

Giancarlo Succi Tullio Vernazza [1] in their work proposes a tool for news extraction, integration, and presentation. Retrieving up-to-date news from many website is rather than simple because traditional search engines are not adequate to support continuous sites updates. The focus is on news extraction and integration problems. They propose an integrated service which includes three main components a News Retriever, a News Repository and a Data Provider. News Retrieval process includes two main steps: news extraction and news integration. News extractor downloads news webpages, extracts data and produces an XML (eXtensible Markup Language) document that can be processed using standard processing tools such as XSLT (eXtensible Stylesheet Language Transformations) to the document to extract relevant data that the news integrator collects as XML documents. News integrator queries XML document collected from the extractor to build a comprehensive document containing all the news for a selected category. At the end of these two processes (news extractor and integrator) the system produces a news repository. Then the data provider retrieves data from news repository and sends them to clients.

Andrea Addis, Giuliano Armano, Francesco Mascia, and Eloisa Vargiu [4] propose a multiagent system for IR. The system extracts the news and articles from the online newspapers, classifies them using hierarchical text categorization also provides the suitable feedback mechanism to the end users to select non-relevant news. In the proposed system once the news are extracted, all the information is suitably encoded to facilitate the text categorization task. To this end, all non-informative words such as prepositions, conjunctions, pronouns and very common verbs are removed using a stop word list. After that, a standard stemming algorithm removes the most common morphological and inflectional suffixes. Then, for each category of the taxonomy, feature selection, based on the information-gain heuristics, has been adopted to reduce the dimensionality of the feature space.

Alexandr Zharikov, Konstantin Kristalovsky, Vasiliy Pivovarov [2] in their work proposed a Natural Language Processing system which retrieve the information about person, organizations or other text objects from an article. The system searches for a query given and present the result to user. Here queries are preprocessed which include the steps chunking, tokenization and morphology marking and querying semantic dictionaries. Then cluster the results and display the essence of list to user.

Categorization of News is an essential step this makes easy to retrieve particular topic. Harmandeep Kaur, Sheenam Malhotra [5] in their work present an algorithm which can classify the inner structures of the simple news clusters. Divide each category into sub categories (ex: sports can be sub categorize to cricket, football, etc.). Some algorithms like K-mean, CART, SVM and HMM helps to classify the clusters into sub clusters, HMM (Hidden Markov Model) is used for text extraction. When we search any newspaper the source code will be displayed it is not in proper text form. Some html tags are also including in this source code. HMM remove all these tags from source code. When html tags are removed then empty space is shown in the place of tags. Then next step of HMM is to remove this empty space. With this we obtain text in proper form. This text is used further as an input in SVM for classification .Then SVM is Support Vector Machine, It is a binary classifier used for text classification. Positive data is representing as 1 and negative data represent as 0. It is used to distinguish the keywords. K mean is used to create the clusters. It groups the data into K clusters. K mean is to minimize the Euclidean distance between data points. It creates the clusters of different categories like clustering for sports is performed as hockey, football, cricket, etc. CART is classification and regression tree. It is used to set the counter with higher value. It creates a hierarchy for the classification of news [6]. With this classification method we can retrieve particular data quickly. Newspaper pages are generally formed from several independent articles which are scattered throughout the page in columns. In the analysis of a newspaper page an important step is the clustering of various text blocks into logical units i.e. into articles. The quantity of documents that needs to be converted into digital format is thus increasing, creating the need for systems capable to extracting knowledge and 'understanding' documents automatically [8].

Most of the news search engines rank the search results based on the relevance of the content of the article to the query, and based on the date when the article was written, preferring newer articles. Lorand Dali, Blaz Fortuna, Jan Rupnik [7] propose personalized ranking model based on click through logs of a large news site. Each search made is recorded by the system like user, query, time and clicked result with this the assumption is made that the article represented by the clicked search result is more relevant than every article which has a higher rank in the search results. Most of relevant result is the most number of users click on result. The model developed taken into account not only document content and metadata but also demographic feature of the user and finds the pair wise relevance from the search logs.

3. Research Challenges of Information Retrieval System for Online Newspapers

Information Retrieval on the Web has always been different and difficult task as compared with a classical information retrieval system (Library System). To explain the difference between classical information retrieval and information retrieval on the Web we compare the two.

We first discuss the differences in the documents.

- **Hypertext**: Documents present on the web are different from general text-only documents because of the presence of hyperlinks. It is estimated that there are roughly 10 hyperlinks present per document.

- **Heterogeneity of document**: The contents present on a web page are heterogeneous in nature i.e., in addition to text they might contain other multimedia contents like audio, video and images.

- **Duplication**: On the Web, over 20% of the documents present are either near or exact duplicates of other documents and this estimation has not included the semantic duplicates yet.

- **Lack of stability**: Web pages lack stability in the sense that the contents of Web pages are modified frequently.

4. Conclusion

As a conclusion this paper explores the retrieval of relevant information making available to various applications and end users. This paper explains the retrieval of related information from different online newspapers for a specific topic and integrating the retrieved data and presenting to end users. So Data Mining provides all the facilities to create an efficient tool for Information Retrieval System. It is observed that clustering technique helps us to group the retrieved news into a list of meaningful categories and Text Classification methods to retrieve news efficiently and quickly.

====

Smartphones and tablets are powerful and popular. More than thousand new mobile apps hitting the market every day. In this fast-moving technological era, is security keeping up? Apps and mobile devices often rely on consumer data — including contact information, photos, and location to name a few — and can be vulnerable to digital snoops, data breaches, and real-world thieves.

The recent security reports from around the world has expressed their serious concern about mobile device apps. According to malware analyst, while there have been no major targeted attacks on mobile devices – as has been the case with desktop platforms in recent years – it is clear that cybercriminals are focusing their attention on smartphones. It's also clear that the cybercriminals are using social networks to get an 'in' on to users' smartphones.

Google, has done a lot to make Android more secure than earlier versions of the OS, but there is much more to be done. Users should also consider using encryption for their data, and only store the data that they really need to access on the smartphone or tablet itself. In this paper we are discussing some tips and measures which can improve the security of mobile apps development.

5. References

[1] Alberto Sillitti. Marco Scotto. Giancarlo Succi ”News Miner: a Tool for Information Retrieval”

[2] Alexandr Zharikov, Konstantin Kristalovsky, Vasiliy Pivovarov “Information Retrieval System For News Articles In Russian”. 2011

[3] Arun K Pujari, “Data Mining Techniques”, published by Universities Press (India) Pvt. Ltd. 2008

[4] Andrea Addis, Giuliano Armano, Francesco Mascia, Eloisa Vargiu: “News Retrieval through a MultiAgent System”. WOA 2007: 48-54

[5] Harmandeep Kaur, Sheenam Malhotra “Inner Classification of Clusters for Online News “ IJCST - Volume1 Issue1, Jul-Aug 2013 ISSN:2347-8578

[6] Krishnalal G ,S Babu Rengarajan, Vallioor, K G Srinivasagan “A New Text Mining Approach Based on HMM-SVM for Web News Classification” 2010 International Journal of Computer Applications (0975 - 8887) Volume 1 – No. 19

[7] Lorand Dali, Bla_z Fortuna, Jan Rupnik “Learning to Rank for Personalized News Article Retrieval” JMLR: Workshop and Conference Proceedings 11 (2010)

[8] Prof. A.D. Thakare, ,Nikita Muthiyan Deepali Nangde, Dipti Patil, Minal Patil "Clustering Of News Articles to Extract Hidden Knowledge" IJETAE (ISSN 2250-2459, Vol2, Issue 11, November 2012)

[9] Ricardo Baeza _Yates, "Modern Information Retrieval" published in the year 2010

[10] Venkat n. Gudivada, vijay v. Raghavan, williami.Grosky, rajesh kasanagottu "Information Retrieval on the World Wide Web" Journal IEEE Internet Computing Volume 1 Issue 5, September 1997 Page58 .