Review of Various Techniques of Automatic Metadata Extraction from Digital Documents

[1]Deval S. Goda & [2]Prof. Dulari A. Bosamiya
[1]ME student, Department of Information Technology, SSEC, Bhavnagar, Gujarat, India
[2]Assistant Professor, Department of Information Technology, SSEC, Bhavnagar, Gujarat,India

***Abstract*— Metadata Extraction is one of the leading research fields in Information Retrieval. Metadata helps providing access to information resources. It makes the discovery and organization of resources substantially easy. The manual collection of such metadata is a tiresome, at times inaccurate and time consuming process. Moreover, the fastest rate at which digital contents are produced will ultimately make it impossible to rely on manual methods. Some automated techniques has to be there for the above said metadata gaining.**

This paper describes some of the available techniques like Metadata extraction with cue model or by TF*PDF algorithm or with the help of support vector machines. One of them is the extraction using a Keyphrase Extraction tool (KET), while the other automatically extracts keywords from web sources.

***Index Terms*— Metadata, Extraction, Information Retrieval**

I. Introduction

Information retrieval formerly used to be a doing that only some people engaged in: for example library officials, and similar professional searchers. Now the era is changing and lacs of people are engaged in information retrieval every day when they use a search engine or search their needs. Information retrieval is fast becoming the foremost form of information access, over theold-style database searching.

Metadata is a data that describe other data. Metadata describes an information source, or helps provide admittance to an information store. A set of such metadata elements may describe one or many information resources. For example, a library catalogue record is a collection of metadata elements, linked to the book or other item in the library collection through the call number. Information stored in the "META" field of an HTML Web page is metadata, linked with the information resource by being within it.

Information Retrieval

The meaning of the term *information retrieval* can be very wide. Just taking a cellphone out of your pocket to search for a contact number is a form of information retrieval. However, as a theoretical field of study, *information retrieval* can be defined as below:

Information retrieval (IR) is finding needed data (usually documents) of an unstructured nature (usually text) that fulfills an information requirement from within large collections (usually stored on computers)

Why is Information Retrieval needed?

Information retrieval is used to retrieve relevant information resources. Volume of information is nowadays becoming grander. Thus, data of required field appears frequently mixed in with other pieces of information that are not of concern. Because the volume of data is now much larger and generally poorly organized, finding useful or desired value might be rather problematic. Actually even with the latest search engines that take profit of link analysis to recognize popular sources of relevant information, the case has been the same.

It is very commonly observed that internet users have poor specifications of their data needs in general. There can be several reasons like if they are in a hurry, or if they do not understand well theprocess of searching, Web users often lay down, very little queries with less or almost nullbackground Information associated with them. They also are the only one at that moment of time to choose what is related and what is not. The Combination of these two factors,suggest that the interaction with users is a crucial step if the precision of the results can be improved. In other words, to improve the user information search we have to ask for more information from him.

IR can also cover other kinds of data and information difficulties more than that given in the fundamental definition above. The term “unstructured data” refers to data which does not have clear, semantically obvious, easy-for-a-computer structure.

Architecture of Information Retrieval system

The architecture of a typical data mining system may have the following major components.

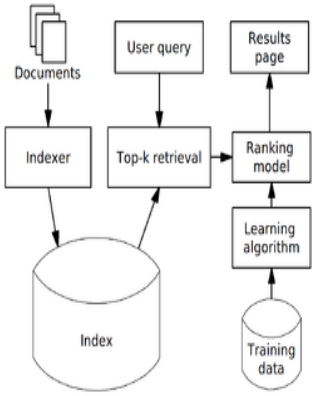

Figure: Architecture of Information Retrieval System [9]

II. Metadata

Metadata is data about data. It helps describing an information resource. A group or of such metadata elements may thoroughly describe one or many information resources.

Whether in the traditional context or in the Internet context, the main aim of metadata is to simplify and improve the extraction of information. At library school, we learnt to measure information retrieval in terms of recall and precision. If we miss a lot of relevant information, we have poor recall. If we get flooded by a lot of irrelevant information, we have poor precision. In certain circumstances (such as searches for patents) very high recall is essential. However, in most circumstances, searchers would be content with a small number of relevant documents, and would be willing to scan through a few dozen citations to identify them. Recall and precision factors of 10-20% are often acceptable for most purposes.

However, our own experiences with Web search engines frequently involve precision factors of much less than even a single percent. For example, a search of the World Wide Web using the search engine ANZWERS on the abbreviation "IETF" (which stands for Internet Engineering Task Force) retrieved 996,354 matches in the year 1998[9]. Every Web page which mentioned the IETF in a related way was recovered by this search.

This example illustrates that search engines can return a lot of unnecessary details as they have no way (or very few means) of differentiating between important and accidental words in documents. If we could work upon the words which are used as major terms, we could achieve a vast improvement in precision. Metadata can be used to achieve this by detecting just the major ideas of the information resource.

If we could target searches onto words or phrases that identify their correct role, we would also improve precision. For example, we could retrieve just those resources where "Global" is related to something showing vastness, without retrieving resources about global warming or environmental issues. Metadata can be used to achieve this by identifying the different features of the information resource: the author, subject, headers, publication companies etc.

There is also a requirement to facilitate search recall - that is, to retrieve documents that would otherwise be missed. For example, websites contain images, databanks, and word or pdf documents along with HTML texts. Metadata can support retrieval of these resources by categorizing them, thus making sure that they are not overlooked by harvesting engines.

Recall can also be improved due to other factors. For example, as per the knowledge available most of the methods of metadata extraction do not index every page on a site, but generally only the top two or three hierarchical stages. Thus, these engines may miss some of the important information which, on larger and more composite sites, may be situated in lower levels of the hierarchy. A better reaping process would gather metadata from a collection created locally from a complete coverage of the local website. The data in this storehouse could then be obtained regularly by the harvesting engine.

Some search (or harvesting) engines do now take account, to some extent, of metadata stored within HTML documents, within the META field.

Performance & Correctness Measure

Many different methods for evaluating the performance of information retrieval systems have been developed. These measures require a gathering of documents and a query. All common processes described here already assume a fact that either every document found is relevant or non-relevant to a particular query. In practice queries may not be properly formatted and there may be different aspects of relevancy.

- **Precision** is the probability that a randomly selected retrieved document is appropriate.[14]
- **Recall** is the probability that a randomly selected relevant document is retrieved in a search.[14]
- **F-measure** is measure that combines precision and recall is the harmonic mean of precision and recall.

  F= 2* ((precision* recall)/ (precision+ recall))

III. Related work

(1)Automatic Document Metadata Extraction using Support Vector Machines
This discusses a machine learning method for automatic metadata extraction. They extend Meta tags from document and mapping it with the Dublin Core 15 Metadata Element [8] Extend the SVM into multi-class classifiers in the "One class versus all others" approaches [1]. Word clustering method is used that contains collection of similar words and use the cluster as a feature. They extracted features based on word-specific & line-specific. Then use that dataset to extract the metadata from documents.
The result can be compared with that of Hidden Markov Model (HMM) as below: (A-Accuracy, P-Precision, R-Recall)

Class	HMM(A)	SVM(A)	SVM(P)	SVM(R)
Title	98.3	98.9	94.1	99.1
Author	93.2	99.3	96.1	98.4
Affiliation	89.4	98.1	92.2	95.4
Address	84.1	99.1	94.9	94.5
Note	84.6	95.5	88.9	75.5
Email	86.9	99.6	90.8	92.7
Date	93.0	99.7	84.0	97.5
Abstract	98.4	97.5	91.1	96.6
Phone	94.9	99.9	93.8	91.0
Keyword	98.5	99.2	96.9	81.5
Web	41.7	99.9	79.5	96.9
Degree	81.2	99.5	80.5	62.2
Pubnum	64.2	99.9	92.2	86.3

(2)Automatic metadata extraction & classification of spreadsheet Documents based on layout similarities
The objective of this is to propose an innovative method that automatically performs metadata extraction and classification on the spread sheets having layout similar to that of a given sample spread sheet whose metadata is previously defined [2]. Metadata classification is based on document types (e.g. purchase order, sales report etc.) and data context (e.g. customer name, order date etc.) XML schema is used to define the classes of metadata of a document type.

- ***s-metadata*** *-stored in a worksheet and* is used to compare with inputted search keywords,
- ***f-metadata*** *- presents field name of the s-metadata,*
- ***p-metadata*** *-presenting file properties (e.g.* file creation date and time etc.) of the spreadsheet.

  Xml Schema is prepared from S-metadata.
  The search index is implemented into PostgreSQL. Because Xml Support PostgrSQL.

(3)An Automatic Online News Topic Keyphrase Extraction System
How to extract key phrases from news topics online automatically? In this paper the state-of-the art TDT techniques are used to organize news pages from a lot of news Websites into topics. An aging theory is added in the TDT process. Unlike previous methods, their system firstly extracts keyword candidates from single news stories, filters them with topic information and then combines them into phrase candidates using position information [3]. Finally, the phrases are ranked and top ones are selected as topic key phrases.
The evaluation can be shown as under:

Strict Evaluation			
Key terms	Precision	Recall	F-score
10	0.43	0.60	0.51
15	0.38	0.75	0.50
20	0.34	0.87	0.49
Lenient Evaluation			
Key terms	Precision	Recall	F-score
10	0.55	0.74	0.64
15	0.44	0.87	0.59
20	0.38	0.97	0.55

(4)KeyPhrase Extraction tool (KET) for semantic metadata annotation of Learning Materials

This paper key-terms extraction based on normalized term frequency, vector of formatting feature and position of the occurrence of the text in the document is proposed and implemented. Here Extract Noun Phrases used OpenNLP Part of Creating a formatting a feature vector then normalizing candidate noun phrase frequency [4]. Then find the Position of word in documents. At last ranking a candidate key phrase

(5)Automated Metadata Extraction from web sources

This paper discusses the application of web wrapping technology in extracting metadata from web sources. This capability has been incorporated into a software tool known as Dynamic Dublin Core/Resource Description Framework Metadata Editor (DDC/RDF-Editor) which supports metadata development and management for resources in the World Wide Web [5]. One key feature of the editor is the ability to automatically extract relevant values of metadata elements from the web sources in question according to the Dublin Core (DC) Metadata Standard and represent it in Resource Description Framework (RDF) language.

*(6)The HOT key phrase Extraction based on TF*PDF*

This paper discussed to extract the key phrase based on modified TF*PDF method. First extracted term extract & then key phrase extracted. First set of term will be created from that set key phrase term will be extracted by examining two term from set which occur consecutively in statement and make a set of key phrase. Based on TF*PDF calculation method hot key phrases are chosen from key phrase sets. This is as shown below

$$W(j) = \sum_{i=1}^{c=|C|} |Fjc| \exp\left(\frac{njc}{Nc}\right)$$

$$| Fjc |= \frac{Fjc}{\sqrt{\sum_{k=1}^{k=k} Fkc}}$$

$$Ps(j) = \begin{cases} 3 & j \in \text{the title of kth document in channel i} \\ 1 & j \notin \text{the title of kth document in channel i} \end{cases}$$

$$Pw(j) = \sum_{i=1}^{|C|} \sum_{k=1}^{|Nc|} psik(j)/Njc$$

$$Weight(j) = W(j) * pw(j)$$

Here, weight(j) is a combination of w(j) (weight of metadata(j) based on frequency) and pw(j) (weight of metadata(j) based on its position in the document).

(7)Metadata Extraction with Cue Model

The proposed new technique described in this paper extracts metadata automatically from document based on a combination of a few programming features that recognize parts of speech, cue words, line position, relative position and symbols to identify the metadata in a particular journal article without using pre-set templates. This technique is easier and time saving than the existing traditional techniques [8].

IV Conclusion and future work

Until now most of metadata extraction system are based on statistical significant of term in the document and its position. Statistical method is useful to extract metadata but have some shortcomings.

-The statistical approach has a weak point that it fails to express vital linguistic features such as phrases.

-The important phrases are not always from the title or the header part. So if we rely on position based methods, the important ones could be missed.

-They require initial training dataset to extract out metadata from documents.

In order to overcome the above shortcomings, we can use NLP (Natural Language Processing) along with the above said statistical methods.

REFERENCES

[1] Hui Han, C. Lee, Eren Manavoglu, Hongyuan Zha, Zhenyue Zhang, Edward A. Fox "Automatic Document Metadata Extraction using Support Vector Machines" In **IEEE** conference in 2003.

[2] Somchai Chatvienchai" Automatic metadata extraction & classification of spread sheet Documents based on layout similarities" In 10th IEEE conference in 2005.

[3] Canhui Wang, Min Zhang, LiyunRu, Shaoping Ma "An Automatic Online News Topic Keyphrase Extraction System" In IEEE conference in 2006

[4] Dr.Jyoti Pareek, Sonal Jain "KeyPhrase Extraction tool (KET) for semantic metadata annotation of Learning Materials" In IEEE conference in 2009

[5] Nor Adnan Yahaya, RosizaBuang "Automated Metadata Extraction from web sources" In IEEE conference 2006.

[6] YAN Gao Jin Liu, Peixun Ma "The HOT keyphrase Extraction based on TF*PDF" In IEEE conference in 2011

[7] Anirvana Mishra, Gaurav Singh "Improving Key phrase extraction by using Document topic information" IN IEEE cinferernce in2012

[8] Wan Malini Wan Isa, Jamaliah Abdul Hamid, Hamidah Ibrahim, Rusli Abdullah, Mohd. HasanSelamat, MuhamadTaufik Abdullah and NurulAmelinaNasharuddin "Metadata Extraction with Cue Model"

[9] http://en.wikipedia.org/wiki/Information_retrieval

[10] http://www.nla.gov.au/openpublish/index.php/nlasp/article/view/1019/1289

[11] http://nlp.stanford.edu/IR-book/html/htmledition/components-of-an-information-retrieval-system-1.html

[12] http://dublincore.org/documents/usageguide/elements.shtml

[13] http://www.nist.gov/speech/tests/tdt/

[14] http://en.wikipedia.org/wiki/precisionandrecall

[15] Bun, K. K., & Ishizuka, M. (2002). Topic extraction using TF * PDF algorithm. In 3rd International conference on Web information systems engineering, (WISE) .pp. 7382