Automatic Extraction of Proper Names and Keywords from HTML Format Paper

Aye Nyein San, Myint Thu Zar Tun
Computer University, Maubin
ayenyeinsan33@gmail.com, myintthuzartun@gmail.com

Abstract

This paper proposes a method to extract keyword from Web Pages (HTML *format paper) automatically. The automatic keyword extraction technique is very important part for the document and web understanding on the Internet. Traditionally, librarians have used the keywords as took for management of information resources and their effective retrieval. This system is based on the premise that names of persons, corporate bodies and keywords present in a text are important in terms their value as search keys for a document. In this paper, we presents methodologies that have been developed and are under the evaluation for the automatic identification and extraction of Name of Persons, Names of Corporate bodies and keywords from web resources.*

Keywords: automatic keyword extraction, librarians, names of persons, name of corporate bodies, keywords, web resources

1. Introduction

Review and comprehension of existing research is fundamental to the ongoing process of conducting research; however, the ever increasing volume of research papers makes accomplishing this task increasingly more difficult. To mitigate this problem of information overload, a form of knowledge reduction may be necessary. It is clear that merit behind extraction of proper names in running text.

The identification of proper names and keywords in written or oral documents is an important task in natural language processing. Proper names constitute a significant part of the text. They account for approximately one third of noun groups and half the words used in proper names do not belong to the French vocabulary (e.g., family names, names of locations, foreign words). In addition, the number of words used in constructing proper names is potentially infinite. [3]

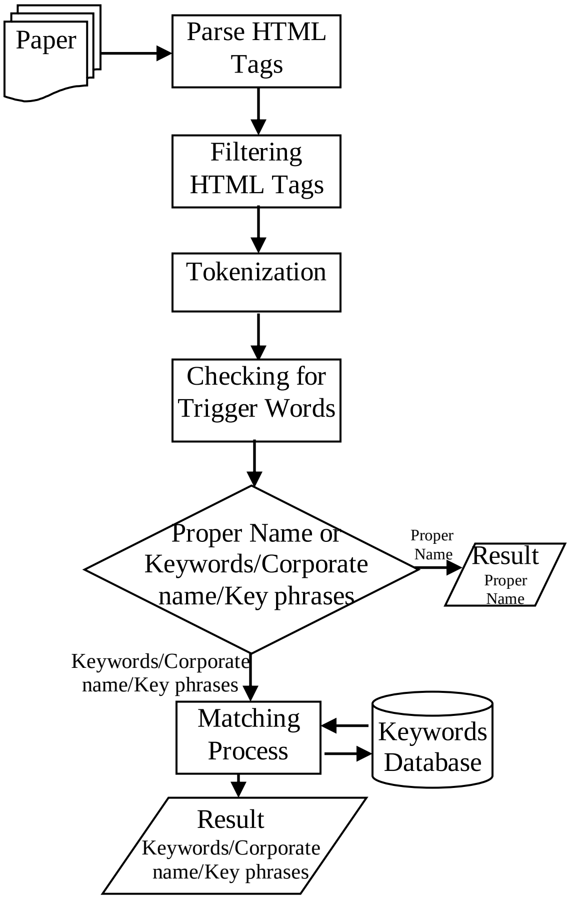
The issue of proper name identification is applied in conferences' papers; in particularly; parallel and soft computing (PSC). If this system use in any paper format, initially, we arrange PSC format paper. After that, this system also performs and analyses these HTML format papers. In these conferences, the first task to achieve is to identify named entities: proper names and also corporate name and paper reference expressions. Corporate name and paper reference expressions are keywords which give a brief summary of a reference papers' contents. With keywords, people can quickly find what they are most interested in and read them carefully. That will save us a lot of time. In addition, these keywords are also useful to the research information retrieval, text clustering and topic search. Manually indexing keywords will cost highly. Thus, automatically indexing keywords from text is of great interests.

Moreover, these research papers can be described in length as numerical numbers of paragraphs and words returned by a query, as the basis for search indexes, as a way of browsing a collection, and as a keyword extraction technique. This task is generally viewed as being generic, in the sense that all texts use such expressions and their expressions and their identification seems a priori independent of the discourse domain or textual genre. This type of text respects strict writing guidelines which facilitates the identification task. For example, sequences like *Mr.* or *Ms.* precedes proper names rather systematically. However, these strategies are insufficient to analyze other types of texts such as a corpus HTML documents because writing guidelines are either different or are much less strict. With the explosion of documents in HTML format, it is precisely these types of documents that need to be processed automatically.

Several methods have been proposed for extracting keywords from text. In this paper, aiming at the characteristics of paper-oriented articles, resources and techniques of current situation, we will introduce a simple procedure to extract keywords and proper names from paper. Section 2 will describe the architecture of the whole system. In Section 3, we will introduce every module in detail of extraction process, including obtaining candidate keywords, how to filter out the meaningless items and how to score possible keyword candidates according to their feature values. In Section 4, performance of the system will be given and analyzed. At last, we will end with the conclusion.

2.System Description

The architecture of our system is shown in Figure 1. We use relaxation strategies to do keyword extraction. In other word, we try to find all the possible keywords first (Phase 1). Then we refine these keywords by some rules (Phase 2).



Paper
Parse HTML Tags
Filtering HTML Tags
Tokenization
Checking for Trigger Words
Proper Name or Keywords/Corporate name/Key phrases
Proper Name
Result
Proper Name
Keywords/Corporate name/Key phrases
Matching Process
Keywords Database
Result
Keywords/Corporate name/Key phrases

Figure 1. The System Architecture of Automatic Extraction System

Phase 1. Possible keywords extraction..

First, the words which can be found in the dictionary are defined as keywords. Then we develop some methods to find the following keywords.

(i)HTML tags: We can extract keywords by some HTML language tags and Hyper-Link, such as, “<head>”,“<title>”, “<H1>”, “<H2>”,…, “<H6>”, “<B>”,“<p>” and so on. The words tagged by these tags are defined as keywords. For instance, <head>Implementation of Text File Security System Using IDEA(International Data Encryption Algorithm</head>,<p>Abstract</p>.

(ii)Proper Name: Most of proper name consists of one word surname and one to two characters first name. We can find family name to the end of comma notation then add one to two characters with dot after family name to form the possible proper name. On the other hand, one or more words with initial upper case letters and further followed by a single uppercase letter followed by a dot or one or more single upper case letters each followed by a dot are also indicators of personal names e.g., Daemen J. and Rijmen V.,

(iii)Corporate Words: Firstly, we find the words corresponding to the company or agency, such as "Co.", "University", "College", "School", etc. Then we also search two to four characters before these words by using keywords database to form the Corporate Words. Prepositions are widely used in corporate names to link legend words with other words (e.g., National Chi-Nan University); this has been exploited to identify complete corporate names.

(iv)New words or Abbreviation words: We use statistical method to find the new words and abbreviation words by using keywords database.

(v)Counting Paragraph or Words: HTML document paper can be calculated the number of paragraphs. We also collect all the words in this document paper to be the number of words.

Phase 2. Keywords Refinement

Since we use the relaxation strategies, there will be too many keywords obtained in phase 1. In order to discard the meaningless keywords, the following strategies are used to refine the keywords.

(i) Parse HTML tags
(ii) Filtering HTML tags
(iii) Tokenization
(iv) Checking for Trigger words
(v) Counting paragraphs/words

(I)Parse HTML Tags and unformatted characters: Any HTML file will necessarily carry tags beginning with angular brackets [<]. These tags did present certain problems. As a first step, therefore, the program remarks all HTML tags and unformatted character like (){}[] etc., from the downloaded or upload files before executing the rest of the program. If a text file is used instead of HTML file this module automatically precedes to the next step of filtering HTML tags. The module does not support other file formats such as PDF, ps, etc. Such files need to be converted to HTML format before applying this model.

(II)Filtering HTML tags: After remarking HTML tags, the program remove or filter the HTML tags. This consists of filtering the HTML files to a format suitable text file and the tokenization process follows as a next step.

(III)Tokenization: Tokenization is the third step. In this process each line of the input text is broken into words and all sequences of capitalized tokens (or words) are collected and stored in a Temp File for further processing. This process is repeated until the entire input text is analyzed.

(IV)Checking for Trigger words: A set of commonly used trigger words (also called as legend words) is stored in three text files called legend (personal names), org names (corporate names) and keywords in keywords database. The system inserts or deletes any data concern with corporate names and keywords before or after processing in the keyword database. Each and every capitalized token extracted is compared with trigger words already stored in the concerning file. Any initial letter that precedes the name, one or two characters with dot are located after name, the comma will be situated to the end of the name and then these words are transferred to Name file. Also an intervening semicolon or colon is used to identify the different names. Such names extracted are also transferred to the Name file. After that corporate name and keywords are matched trigger words from the keywords database, and found these words in keywords database, corporate name are transferred to org file and keywords are also stored in keywords file.

(V)Counting paragraphs/words: When a text file contains more than one paragraph, it is in fact a reckoning structure that makes to be count. In text file, there are also computing the number of words during extracting process. Moreover, each word/phrase extracted from the text file. After that these extracting words are counted when a match is found in keywords database. Such these same words are described to the interface.

3. Extraction Process

Keywords are usually chosen manually. In many academic contexts, the readers assign keywords to documents they have read. However, the great majority of documents come with keywords, and assigning them manually is a tedious process that requires knowledge of the subject matter. Automatic extraction techniques are potentially of great benefit.

There are two fundamentally different approaches to the problem of automatically generating keywords for a text file: name extraction process and keywords/key phrases extraction process.

3.1 Name Extraction Process

There are two kinds of named entities: person name and corporate name. The first are those which have the above rules of composition in Section 2, mainly, initial capital letter with words followed by one or two capital letters with dot at the end of comma. The can be recognized with statistical and rule-based methods combined. These person names in paper file are composed of family names and first names whose lengths are respectively one or two capital letters. Furthermore, there is a relatively stable set of family names and one or two characters followed by comma which often provide the anchor to search a person name. [4]

3.2 Keywords/ Key Phrases Extraction Process

In this process, there are two types of stages: keyword assignment and keyword extraction. This process executes cooperate name and identification keywords. Both use machine learning methods, and require for training purposes a set of paper files with keywords already attached. [1]

Keyword assignment seeks to select the words from a controlled vocabulary that best describe the paper. The training data as shown in Table 1 associates with keyword database which builds a classifier for each keyword. The only keywords that can be assigned are ones that have already been seen in the training data. In keywords database, the elements representing the subject of resource usually the form of a set of data fields that may include keywords, descriptors, subject headings, abstract and classification codes, etc.

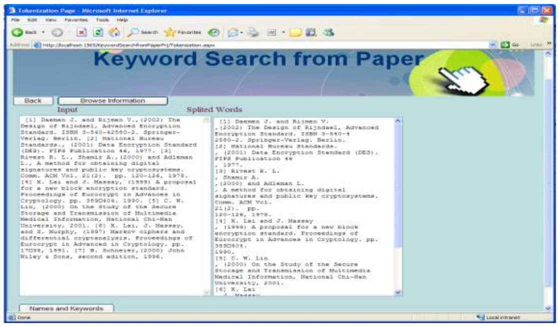
Corporate Name	Keywords/KeyPhrases
ACM	Comm
ISBN	cryptanalysis
National Chi-Nan University	Cryptology
Georgia Institute University of Chicago	Data DES

Table 1. Training Data Set of Keywords Database

Keyword extraction does not use a controlled vocabulary, but instead chooses keywords from the text itself. It employs lexical and information retrieval techniques to extract words from the text file that are likely to characterize it. In the approach, the training data is used to tune the parameters of the extraction process. [2]

4. Performance of the System

In this system, we select any paper in HTML file from PSC format. During the selection, we choice the download or upload files. After selection process, we parse this file for modifying HTML code. The next step is filter process. This process removes the HTML source code and then tokenization process is followed to split words as shown in Figure 2.

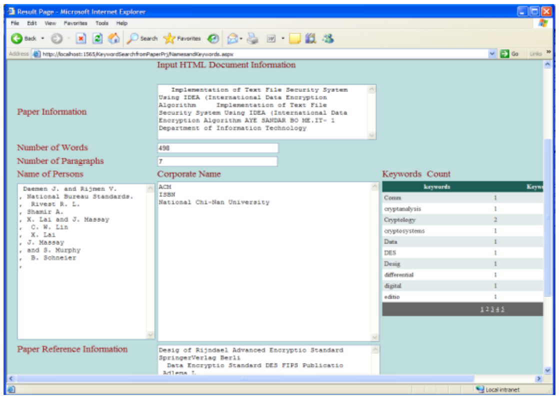

Tokenization Page - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://localhost:1565/KeywordSearchfromPaperPrj/Tokenization.aspx
Go Links
Keyword Search from Paper
Back
Browse Information
Input
Splited Words
Names and Keywords
Done
Local intranet

```
 [1] Daemen J. and Rijmen V.,(2002) The
Design of Rijndael, Advanced Encryption
Standard. ISBN 3-540-42580-2. Springer-
Verlag. Berlin. [2] National Bureau
Standards., (2001) Data Encryption Standard
(DES). FIPS Publication 46, 1977. [3]
Rivest R. L., Shamir A.,(2000) and Adleman
L., A method for obtaining digital
signatures and public key cryptosystems.
Comm. ACM Vol. 21(2). pp. 120-126, 1978.
[4] X. Lai and J. Massey, (1999) A proposal
for a new block encryption standard.
Proceedings of Eurocrypt in Advances in
Cryptology. pp. 389D404. 1990. [5] C. W.
Lin, (2000) On the Study of the Secure
Storage and Transmission of Multimedia
Medical Information, National Chi-Nan
University, 2001. [6] X. Lai. J. Massey,
and S. Murphy, (1997) Markov ciphers and
differential cryptanalysis. Proceedings of
Eurocrypt in Advanced in Cryptology. pp.
17D38, 1991. [7] B. Schneier,(2000) John
Wiley & Sons, second edition, 1996.
```

```
 [1] Daemen J. and Rijmen V.
,(2002) The Design of Rijndael, Advanced
Encryption Standard. ISBN 3-540-4
2580-2. Springer-Verlag. Berlin.
[2] National Bureau Standards.
, (2001) Data Encryption Standard (DES).
FIPS Publication 46
, 1977.
[3] Rivest R. L.
, Shamir A.
,(2000) and Adleman L.
, A method for obtaining digital
signatures and public key cryptosystems.
Comm. ACM Vol.
21(2). pp.
120-126, 1978.
[4] X. Lai and J. Massey
, (1999) A proposal for a new block
encryption standard. Proceedings of
Eurocrypt in Advances in Cryptology. pp.
389D404.
1990.
[5] C. W. Lin
, (2000) On the Study of the Secure
Storage and Transmission of Multimedia
Medical Information, National Chi-Nan
University, 2001.
[6] X. Lai
```

Figure 2. Executing the Tokenization Process for Paper

Finally, there are paper title, paper information (author, location), number of words, number paragraphs, person names, corporate name and keywords by using extraction process as shown in Figure 3. Here, we automatically extracted keywords from them and evaluated the amount of keywords/ key phrases to show the user interface.


Result Page - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://localhost:1565/KeywordSearchfromPaperPrj/NamesandKeywords.aspx
Go Links
Local intranet

Input HTML Document Information

Paper Information

```
   Implementation of Text File Security System
Using IDEA (International Data Encryption
Algorithm     Implementation of Text File
Security System Using IDEA (International Data
Encryption Algorithm AYE SANDAR BO ME.IT- 1
Department of Information Technology
```

Number of Words: 498

Number of Paragraphs: 7

Name of Persons

```
 Daemen J. and Rijmen V.
, National Bureau Standards.
,  Rivest R. L.
, Shamir A.
, X. Lai and J. Massey
,  C. W. Lin
,  X. Lai
, J. Massey
, and S. Murphy
,  B. Schneier
,
```

Corporate Name

```
ACM
ISBN
National Chi-Nan University
```

Keywords Count

keywords	Keyw
Comm	1
cryptanalysis	1
Cryptology	2
cryptosystems	1
Data	1
DES	1
Desig	1
differential	1
digital	1
editio	1
1 2 3 4 5	

Paper Reference Information

```
Desig of Rijndael Advanced Encryptio Standard
SpringerVerlag Berli
  Data Encryptio Standard DES FIPS Publicatio
 Adlema L
```

Figure 3. Results of Automatically Extraction of Proper Names and Keywords

5. Conclusion

One of the major Challenges facing information professionals today relates to effectives mechanisms for retrieval of information from Web. This system uses reasonably fast and robust heuristics to identify proper names, keywords and extract them from HTML format paper for librarians. These librarians are easy to use, search and retrieve the person name, corporative bodies and keywords by using this system for the library. It is relevant to indicate here that it took auto extract to generate the output and create a link HTML file linking keywords to the source document. An idea of the output, i.e., keywords and key phrases is extracted by automatically extraction process. The program in its present form can be used only with HTML pages in the English Language. The program requires the online availability of a good thesaurus / glossary of terms in the subject domain for its effective functioning. This requires that the glossary must be regularly and frequently updated by addition of new domain terms, deletions and modifications that may be necessary.

References

[1] E.Frank, G.W. Paynter, I.H. Witten, C. Gutwin, and C.G. Nevill-Manning., "Domain-specific keyphrase extraction", Proc. Sixteenth International Joint Conference on Artificial Intelligence, Morgan Kaufmann Publishers, San Francisco, CA, 1999, pp. 668-673.

[2] I.H. Witten, G.W. Paynter, E. Frank, C. Gutwin, and C.G. Nevill-Manning, KEA: "Practical Automatic Keyphrase Extraction", Proc. DL '99, 1999, pp. 254-256

[3] L. Yu-Sheng, W. Chung-Hsien, "Meaningful term extraction and discriminative term selection in text categorization via unknown-word methodology", ACM Transactions on Asian Language Information Processing (TALIP), Vol.1, No.1, March 2002, pp. 34-64.

[4] P.D. Turney, "Learning to Extract Key-phrases from Text", NRC Technical Report ERB-1057, National Research Council, Canada, 1999.