Extraction of Punjabi Keywords For Database of Search Engine

Puneet Wadhawan[1], Sapna Dhiman[2]

[1] Department of Computer Science, M.M.Modi College,

Patiala, Punjab, India

[2] Department of Computer Science, M.M.Modi College

Patiala, Punjab, India

Abstract

Huge amount of information is available on web. To retrieve that information from web we need a special tool i.e search engine. Search engine works very fast. The result of search engine depends upon keywords typed by user. If that keyword is present in database of search engine searching will be done otherwise error message is generated. In this paper, we discussed method to extract of Punjabi keywords from downloaded text which will be stored in database for searching. Unigram and Bi-gram keywords are taken for database.

Keywords: search engine, unigram, bigram

1. Introduction

There is huge amount of data and information is available at web. More over thousands of new documents are created and changed every day across the internet. The amount of information is increases exponentially. So there is a need of special tool which helps to retrieve information from web.

A search engine is a software program that searches for sites based on the keywords and returns a list of the documents where keywords are found.

The search results are usually presented in a list and commonly called hits. The search is fully dependent on keyword which is typed by used i.e user type keyword and search engine displays different web-pages or data for that keyword. But that keyword must be included in database of that search engine. So, the initial stage of any search engine development is to create database based on these keywords. Keywords are the set of significant words and identifying these keywords from large amount of text is also a challenging

Copyright (c) 2012 International Journal of Computer Science Issues. All Rights Reserved.

work[3]. In this paper, we discussed the method to extract of Punjabi keywords from downloaded text. Unigram and Bigram keywords are extracted with their frequency. Unigram is single word while Bigram is combination of two words. The frequency shows how much time that word is found in that particular page.

2. Related Work

Gerard Salton was the father of modern search technology. The first SMART information retrieval system was developed by his team Harvard and Cornell[2]. This system includes some important concepts like vector space model, inverse document frequency, term frequency and relevancy feedback mechanisms. Concept of hypertext was included by Ted Nelson in his project Xenadu. There is very short and brief history of search engine is available at web. The first search engine on web is World Wide Web Worm (wwww) introduced by McBryan in 1994[2]. That search engine was followed by some other academic search engines. After the concept of crawler system many search engines are coming in market. Lycos was the search engine is first depending upon this concept. Lycos was designed by Michale Mauldin in 1994. After this Altavista, Dogpile, Ask jeeves, AlltheWeb, Google, Yahoo, MSN Search, Ask.com, GoodSearch, Live Search etc. are coming on web. But all of them Google search engine is very popular for searching information. The working of different search engines is different but all search engines generally perform three task[1,4]

- They search the Internet -- or select pieces of the Internet -- based on important keywords.
- They keep an index of the keywords they find, and where they find them.
- They allow users to look for keywords or combinations of keywords found in that index.

3. Our Approach To Extract Keywords

Before developing any search engine the main work is to collect and design corpus of that search engine. This corpus helps to decide keywords for search engine. The best source of collecting corpus is internet or direct communication with different persons. Many newspapers, books, Punjabi stories are downloaded from different websites but these web pages include Punjabi and non-Punjabi text. So for our research work, Punjabi text is separated from non-Punjabi text and stored into different files. Unigram and Bigram are extracted from Punjabi text. To extract Unigram and Bigram AKHAR software is used. Punjabi text source files are imported into AKHAR software one by one and it generated target files which contain Unigram data with their frequency, then Bigram data with their frequency and then Trigram data. But for our research work, we only use Unigram and Bigram data. Now from these target files Unigram data and Bigram data are stored into two different files, one for all Unigram data and second for all Bigram data.

Copyright (c) 2012 International Journal of Computer Science Issues. All Rights Reserved.

Example of Unigram

ਖੇਡਾਂ, ਬਿਜ਼ਨਸ, ਮਨੋਰੰਜਨ, ਪੰਜਾਬੀ

Example of Bigram

ਪਰਿਵਾਰ, ਸਿਹਤ , ਪਰਿਵਾਰਾਂ, ਦੇ

The next and important task is to decide keywords from these two files. These keywords help for searching data and web pages from internet. These Unigram and Bigram can't be used as keywords because they contain all common and uncommon words. In our Punjabi language 480 words are most common words and these common words are generally not used as keywords. It means our Unigram and Bigram files should not contain these common words. Two separate algorithms are designed for extracted keywords from both files. Some most common Punjabi words are:

ਦੇ , ਨੂੰ, ਵਿਚ, ਸੀ, ਕੀ, ਹਨ, ਨਹੀਂ

3.1 Extraction of Unigram keywords

Unigram contains single words, hence each string of Unigram file is compared with common words of Punjabi to select Punjabi keywords from them.

P is a text file which stores most common 480 words of Punjabi language, U is text file which stores Unigram with frequency, A is 2D array and U1 is also text file in which final output is stored. Working of algorithms is:

1. File P is open in read mode
2. Read Punjabi string from the file P and stored that string into A
3. Open U in read mode and U1 in write mode
4. Read one by one Unigram string form U and compare it with each 480 strings of A.
5. If match is not found then write that string with its frequency into U1. Read next string from U.
6. If match found, read next string from U and repeat step 4 and 5 for that string till end of file.

The output of this algorithm provides a file which contains Unigram keywords and not containing any common word.

Example

Unigram word	frequency
ਪੰਜਾਬੀ	2
ਪੰਨ੍ਹਾ	1
ਪਰਿਵਾਰ	2
ਪਰਿਵਾਰਾਂ	2
ਉਕਤ	5
ਉਮੀਦਵਾਰ	15

Table 3.1 Shows unigram keyword frequency

3.2 Extraction of Bigram keywords

After completion of Unigram keywords next step is to extract Bigram keywords from Bigram files. In case of Unigram, only one string is taken but in Bigram we have combination of two strings. So we have to compare both strings with common words of Punjabi.

Copyright (c) 2012 International Journal of Computer Science Issues. All Rights Reserved.

P is a text file which stores most common 480 words of Punjabi language, B is text file which stores Bigram with frequency, A is 2D array, w1 and w2 are string variables and B1 is also text file in which final output is stored. Working of algorithms is:

1. File P is open in read mode
2. Read Punjabi string from the file P and stored that string into A
3. Open B in read mode and B1 in write mode
4. Read each Bigram string form B
5. Stored first string of Bigram into w1 and second in w2
6. Compare w1 with each string of A , If match is not found then compare w2 with each string of A.
7. If match is not found for both w1 and w2, write these words with their frequency into B1.
8. If match found, exit from the loop and read next string from B and repeat step 4 to 7 for that string till end of file.

This algorithm produces a target file which contains Bigram keywords.

Example

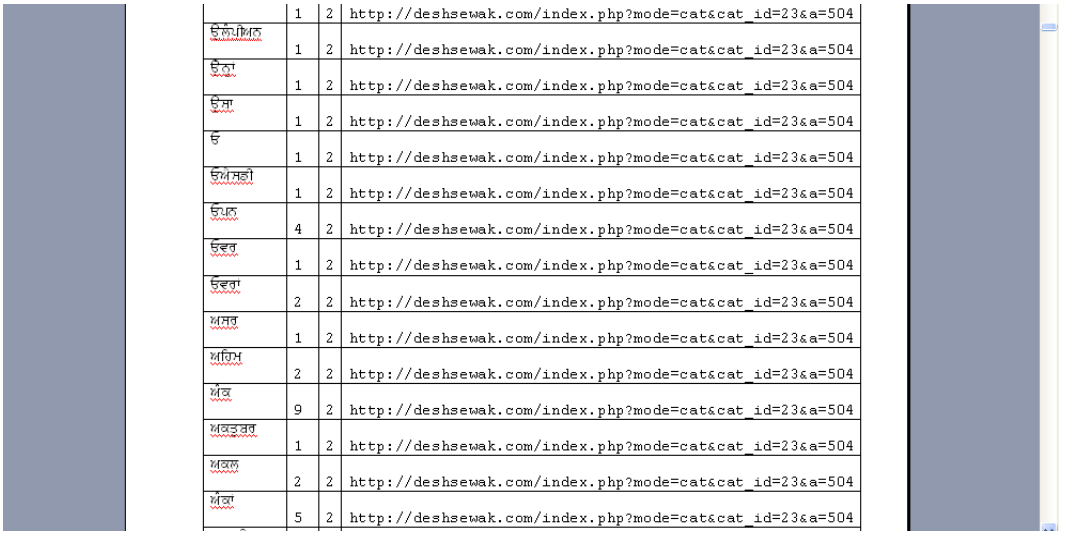
word1	word2	frequency
ਯੂ	ਐਨ	1
ਅੰਗ	ਸੰਗ	3
੍ਆਹ	ਪ੍ਰਨੀਤ	1
ਭਾਰਤੀ	ਸਟੇਟ	2
ਫਰੀਦਕੋਟ	ਸਪਨ	1
ਯਾਦਗਾਰੀ	ਫ਼ਿਲਮ	1
ਆਖ਼ਰੀ	ਯਾਤਰਾ	7
ਆਖ਼ਰੀ	ਵਿਦਾਈ	1

Table 3.2 Shows Bigram keywords with frequency

Now these Unigram and Bigram keywords are used for searching purpose. They are stored into database. When user typed any keyword on user interface, first of all that keyword is searched into database. If match will be found then searching of web-pages is performed by search engine otherwise an error report is generated by system.

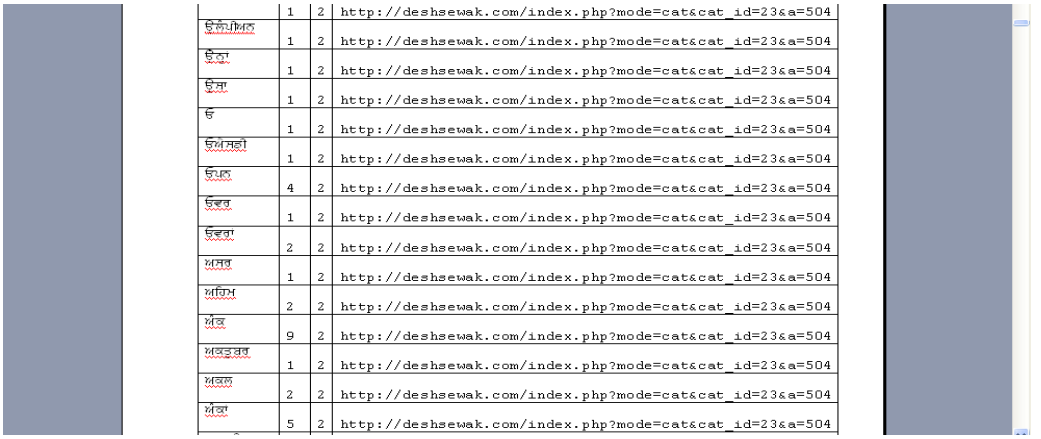
Copyright (c) 2012 International Journal of Computer Science Issues. All Rights Reserved.

4. Outputs Generated by Algorithms



	1	2	http://deshsewak.com/index.php?mode=cat&cat_id=23&a=504
ਉਲੰਪਿਅਨ	1	2	http://deshsewak.com/index.php?mode=cat&cat_id=23&a=504
ਉਨ੍ਹਾ	1	2	http://deshsewak.com/index.php?mode=cat&cat_id=23&a=504
ਉਸ	1	2	http://deshsewak.com/index.php?mode=cat&cat_id=23&a=504
ਓ	1	2	http://deshsewak.com/index.php?mode=cat&cat_id=23&a=504
ਓਅਸਡੀ	1	2	http://deshsewak.com/index.php?mode=cat&cat_id=23&a=504
ਓਪਨ	4	2	http://deshsewak.com/index.php?mode=cat&cat_id=23&a=504
ਓਵਰ	1	2	http://deshsewak.com/index.php?mode=cat&cat_id=23&a=504
ਓਵਰਾ	2	2	http://deshsewak.com/index.php?mode=cat&cat_id=23&a=504
ਅਸਰ	1	2	http://deshsewak.com/index.php?mode=cat&cat_id=23&a=504
ਅਹਿਮ	2	2	http://deshsewak.com/index.php?mode=cat&cat_id=23&a=504
ਅੰਕ	9	2	http://deshsewak.com/index.php?mode=cat&cat_id=23&a=504
ਅਕਤੂਬਰ	1	2	http://deshsewak.com/index.php?mode=cat&cat_id=23&a=504
ਅਕਲ	2	2	http://deshsewak.com/index.php?mode=cat&cat_id=23&a=504
ਅੰਕਾ	5	2	http://deshsewak.com/index.php?mode=cat&cat_id=23&a=504

Fig 4.1 shows Punjabi unigram table have four columns (i.e. word, frequency, filename and url).



	1	2	http://deshsewak.com/index.php?mode=cat&cat_id=23&a=504
ਉਲੰਪਿਅਨ	1	2	http://deshsewak.com/index.php?mode=cat&cat_id=23&a=504
ਉਨ੍ਹਾ	1	2	http://deshsewak.com/index.php?mode=cat&cat_id=23&a=504
ਉਸ	1	2	http://deshsewak.com/index.php?mode=cat&cat_id=23&a=504
ਓ	1	2	http://deshsewak.com/index.php?mode=cat&cat_id=23&a=504
ਓਅਸਡੀ	1	2	http://deshsewak.com/index.php?mode=cat&cat_id=23&a=504
ਓਪਨ	4	2	http://deshsewak.com/index.php?mode=cat&cat_id=23&a=504
ਓਵਰ	1	2	http://deshsewak.com/index.php?mode=cat&cat_id=23&a=504
ਓਵਰਾ	2	2	http://deshsewak.com/index.php?mode=cat&cat_id=23&a=504
ਅਸਰ	1	2	http://deshsewak.com/index.php?mode=cat&cat_id=23&a=504
ਅਹਿਮ	2	2	http://deshsewak.com/index.php?mode=cat&cat_id=23&a=504
ਅੰਕ	9	2	http://deshsewak.com/index.php?mode=cat&cat_id=23&a=504
ਅਕਤੂਬਰ	1	2	http://deshsewak.com/index.php?mode=cat&cat_id=23&a=504
ਅਕਲ	2	2	http://deshsewak.com/index.php?mode=cat&cat_id=23&a=504
ਅੰਕਾ	5	2	http://deshsewak.com/index.php?mode=cat&cat_id=23&a=504

Fig 4.2 shows bigram table have five columns (i.e. word1, word2, frequency, filename and url).

Copyright (c) 2012 International Journal of Computer Science Issues. All Rights Reserved.

4. Conclusion And Future Work

For the research most of the text is downloaded from Punjabi Encoded Websites, which consist Punjabi and non-Punjabi text. These methods produce uncommon Unigram and Bigram keywords for database. But this research work will be extended for n-gram keywords. This research is carried out for Punjabi Language only. In future for other languages, database will be developed.

References

[1]. S. Brin and L. Page, "The anatomy of large scale hypertextual web search engine", in Proceedings of the 7th Internatioanl World Wide Web Conference, Brisbane, Australia, 1998, page no. 107-117.

[2] A. Wall, "History of Search Engines: From 1945 to Google 2007", Search Engine History, 2001, Available at: http://www.searchenginehistory.com/

[3] V. Gupta and G. S. Lehal, "A Survey of Text Mining Techniques and Application"

[4]. A. Arasu, J. Cho, H. Garcia-Molina and S. Raghavan, "Searching the Web", Published by ACM on Inernet Technologies, 2001, Vol. 1, page no. 2-43.

Copyright (c) 2012 International Journal of Computer Science Issues. All Rights Reserved.