Big Data Application in Web Search and Text Mining

Wenxuan Han
Indiana University Bloomingtonn
1150 S Clarizz Blvd
Bloomington, Indiana 47401-4294
wenxhan@iu.edu

ABSTRACT

Because of the rapid development of social media, there are gigantic amount of data generated in every second on the web. And those data could be stored in any forms like text, videos, images or their combinations. The more complicated forms of data, the more space it will take up and will cost more time to read it. Although most of today's personal computers have a very high performance, it is extremely difficult to process and analyze useful text information from those huge amount of unstructured data by using traditional single computer methods without the help of big data tools or text mining techniques. Fortunately, the improvements in big data application are also increasing fast in order to support those difficult works on web search and text mining. This paper first studies the knowledge of web search technique and its data analytic steps, then introduces the link structure with a broad analysis of some web page structures (Hubs and PageRank), and at last, discusses their applications in this field of big data.

KEYWORDS

I523, HID209, Big Data, Social Media, Web Search, Text Mining, PageRank, Hubs

1 INTRODUCTION

In recent years, social media has become more and more popular as a new way of communication and knowledge transfer. People could use it to create, share, exchange information and create their own network. Social media usage has been boosted from 2005 to 2015. Users between 18 and 29 ages are the mainly part of social media users [7]. Today 90% of young adults are active on social media. This proportion was 12% in 2005 [1]. And since the development of mobile products, social media has also been offered a better platform for users to share data faster and more convenient. Thus, this proportion could be keep stable or still increase during the next few years.

Nowadays, a growing number of people prefer to express their opinion and feelings through tweeting, sharing images, commenting on social sites [7]. Since the amount of such data become extremely large, it is significant to extract and analyze useful information through them by using text analysis methods. Therefore, some applications which based on these information have been developed, such as recommendation system and search engine.

However, as the big data began to appear in the website, there are some problems that people must face for web search which include the longer search queries (key words) requirement, support the huge number of searches and multiple languages. And these problems cause the progress of web search and text mining technologies.

Web search is similar to information retrieval (IR) which is used to search for information on the World Wide Web [10]. The information may be a mix of web pages, images, and other types of files. Since web search is applying on web which has the huge amount of data, it has a much larger scale than many IR systems. Although web search is a complex technique, it has the capability to understand how to crawl internet to get and update information.

Text mining (also known as knowledge discovery in text database [4]) is semi-automatic process of discovering information, meaningful contents, topics, word, relations and patterns from a large amount of text data [7], which is also a branch of data mining. The text data could be extracted by web search at first.

2 WEB SEARCH TECHNIQUE

2.1 Key Fundamental Principles

DIKW hierarchical model is the most basic model in the information management, information systems and knowledge management disciplines. Thus, it also used behind web search technique. It contains four main components: data, information, knowledge and wisdom. Since this paper only considers this model in web search area, these four components have the following conception.

- Data: raw web pages or "documents viewed as a bag of words".
- Information: result of query or "documents viewed as a collection of insights".
- Knowledge: result of processing query results by user.
- Wisdom: synthesis of many such actions by a set of users.

Fighure 1 shows the hierarchical framework of DIKW model. It shows a pyramid contructure with wisdom in the top level and data in the bottom level.

[Figure 1 about here.]

2.2 Search Engines

A web search engine is a software system for searching information on the Internet. The search results are generally presented in a line which are often referred to as search engine results pages. And some search engines also haeve the capability to mine data from databases or other open directories. Unlike web directories, which are maintained only by human editors, search engines also maintain web crawling, indexing and searching processes in real-time [10]. Table 1 displays the development of search engines and some searching technologies in recent years.

[Table 1 about here.]

2.3 Boolean and Vector Space Models

After discussed the basic principles and the application of web search, here introduce a model that used to define the search technique. Boolean model and vector model are both retrieval model that can be a description of either the computational process or the human process of retrieval. For a retrieval model, it specifies the details of [6]:

- Document representation.
- Query representation.
- Retrieval function (how to find relevant results).
- Determines a notion of relevance.

In boolean model, keywords are considered to be either present or absent in a document and to provide equal evidence with respect to information needs. Queries are boolean expressions of keywords, which connected by AND ($\wedge$), OR ($\vee$), and NOT ($\neg$), including the use of brackets to indicate scope [6]. Thus, for the output of this model, the result document should be either relevant or not, and could not give partial matches or a ranking. Although this model is easy to understand and offers a clean formalism, it might become extremely complicated for most of web users in big data.

For vector space model, documents and queries are vectors in a high-dimensional space. Assume t distinct terms remain after preprocessing. Each term (i) in a document or query (j) is given a real valued weight w_{ij}. Therefore, both documents and queries are expressed as t-dimensional vectors [6]:

$$d_j = (w_{1j}, w_{2j}, \cdots, w_{tj})$$

There are some patterns to represent term weight. One is the Term Frequency, which assume that important terms have the higher frequency of occurrence in a document. The following equation define the vector space model.

$$tf(t,d) = \begin{cases} 0, & freq(d,t) = 0 \\ 1 + \log freq(d,t), & \text{otherwise} \end{cases}$$

While t refers to terms, and d refers to documents. This model straightforward to map everything to a vector and compare their angles. But it is hard to find a good set of basis vectors, a good weighting scheme for terms and a comparison function.

2.4 Web Crawling

It is not difficult to extract data from web with the help of algorithms. The input of the algorithm could be a list of URL's visited already and a list of new URL's to visit. Then executes the following steps in a loop:

(1) Fetch URL off list and check if done.
(2) If not done, go to web and continue collect.
(3) Hand document to document analyser.
(4) Extract all URL's and add to list of new URL's to visit.

The result could be lots of detail of course. Then after fetching from the web, it should do the following steps:

(1) Convert document from HTML, PDF, Word, . . . to a text document.
(2) Tokenization: remove formatting, punctuation, capitals and convert to common form which makes document become a set of canonical tokens.
(3) Filtration: remove "stop words" (e.g. the, is, a, etc.).
(4) Stemming and Normalization: remove inflections and cope with non trival synonyms.

Then the output are contents in bag of words and final terms are those used to define each dimension of vector space model.

3 WEB DATA (TEXT) MINING

3.1 Web Data Analytics Steps

For the big data which people search from web, it could be very difficult to extract or analyze useful information behind them. Thus, it is necessary to define the following steps to make those data structured or orderly so that people could easily applying other techniques like text mining to analyze them.

(1) Get the digital data from web.
(2) Preprocess data into searchable data like words or positions.
(3) Form Inverted Index in order to map words to documents.
(4) Use algorithm like PageRank to rank relevance of documents.
(5) Apply some technologies (e.g. reverse engineering, preventing reverse engineering, etc.) for web advertising.
(6) Build the structure of the Internet and its people and pages.
(7) Clustering documents into topics.
(8) Might utilize Bayes to convert Mathematics of frequency into Mathematics of belief.

3.2 Link Structure Analysis

Since link structure has the significant impact to Search Engine Rankings, the PageRank flow and the number of pages that get indexed, it became one of the important factors of SEO (Search Engine Optimization) [2].

Link structure explores the connectivity patterns between web pages that contain the useful information and makes the huge of websitefis statistics meaningful. That is to say, mining these big data could help us understand what kind of things that users looking for, what are the hottest categories of a website and which pages are the most popular. Continuous optimization of link structure can eliminate duplicate content and promote popular pages in order to get more pageviews and higher rankings on Search Engine results [2].

An idea of the link structure for web pages is Hubs, which is known as Hubs and Authorities. The concept of this idea is simple: certain web pages served as large directories that were not actually authoritative in the information for users, but have links that led users direct to other authoritative pages [5]. Figure 2 shows the structure of Hubs.

[Figure 2 about here.]

As it shows in this figure, a good hub represented a page that pointed to many other pages and a good authority represented a page that was linked by many different hubs.

After defined the link structure of web pages, it comes to a link analysis algorithm named PageRank used by Google Search to rank websites in their search engine results. It is a way of measuring the importance of website pages. PageRank assigns a numerical weighting to each element of a hyperlinked set of documents with the purpose of "measuring" its relative importance within the set.

The numerical weight that it assigns to any given element E is referred to as the PageRank of E and denoted by $PR(E)$ [9]. The output of PageRank is a probability distribution that a page will be visited by a person who has the same probability to click each link on this page. This probability could be calculated iteratively with each page getting a contribution at each iteration equal to its page rank divided by the sum of links on page:

$$PR(\text{page } i) = \Sigma_{\text{page } j \text{ pointing at } i} \frac{PR(\text{page } j)}{\text{number of pages linked on page } j}$$

For example, a PageRank of 0.5 means there is a 50% chance that a person clicking on a random link will be directed to the document with the 0.5 PageRank [9]. PageRank could be used in ranking academic doctoral programs, recommendation systems and many other searching areas.

3.3 Clustering and Topic Models

After obtained results through a search query, it is important to classify them by groups for the further analysis. Clustering, also known as grouping ducument together, is the responses to a search query which give a group of documents. Suppose documents are the points in a space, the task of clustering is to identify regions. There are several ways to do this task:

- Clustering: Nearby regions of points.
- Support Vector Machine (SVM): Chop space up into parts.
- Gaussian Mixture Models (GMMs): A type of fuzzy clustering.
- K-Nearest Neighbors.

Alternatively, some "hidden meaning" can be determined with a topic model. It used to discover the abstract "topics" that occur in a collection of documents so that people could group documents by those topics. Topic modeling is a frequently used text-mining tool for discovery of hidden semantic structures in a text body [3]. Assume each document is a set of topics and each topic is a bag of words, a topic model aims to find the best set of topics and best set of words in topics through a mathematical framework. That is to say, it allows people to examine a set of documents and discover what the topics might be and what each document's balance of topics is [3].

4 CONCLUSION

The aim of this paper is to demonstrate the core contents and the technique background of web search and text mining in big data area. Since the growth amount of data generate on web everyday cause the traditional computing methods and algorithms inefficiency, it is essential to make innovations in web search aspect. In the recent twenty years, search engines developed quickly and DIKW model, which was known as a popular model used in information system before, has applied in web for building its basic principles as well. As the vector space model appeared, the simple boolean model has been replaced in order to define the search query model more completely. And with the help of web crawling algorithm, multiple types of text data extracted from website have become normalized before mining (analysis) the useful information.

Since webs page could seem as link structure, there must exists some patterns between linked pages. PageRank which found by Google is still widely applied in many different big data systems today, it has the ability to find the most relevant page for the content that the user searches for. After obtained pages of data information, we could utilize clustering to group documents together by topics.

ACKNOWLEDGMENTS

The author would like to thank Professor Gregor von Laszewski and all TAs for providing the resource, tutorials and other related materials to write this paper.

REFERENCES

[1] Perrin A. 2015. Social Networking Usage: 2005-2015. (Octobe 2015).
[2] Bbriniotis. 2016. Link Structure: Analyzing the most important methods. (October 2016). http://www.webseoanalytics.com/blog/link-structure-analyzing-the-most-important-methods/
[3] David Blei. 2012. Probabilistic Topic Models. *Commun. ACM* (2012).
[4] Emir and Almir. 2016. Application of Big Data and Text Mining Methods and Technologies in Modern Business Analyzing Social Networks Data about Traffic Tracking. *IEEE* (October 2016).
[5] Christopher D. Manning and Prabhakar Raghavan. 2008. Introduction to Information Retrieval. *Cambridge University Press* (2008).
[6] R.Mooney, J. Ghosh, and D. Lee. 2017. Boolean and Vector Space Retrieval Models. (2017). http://www.cs.ucsb.edu/~tyang/class/293S17/slides/Topic2IRModels.pdf
[7] Mehmet U. and Secren G. 2016. Text Mining Analysis in Turkish Language Using Big Data Tools. *IEEE Computer Society* (2016).
[8] Wikipedia. 2017. DIKW pyramid. (September 2017). https://en.wikipedia.org/wiki/DIKW_pyramid#cite_note-Rowley-1
[9] Wikipedia. 2017. PageRank. (September 2017). https://en.wikipedia.org/wiki/PageRank
[10] Wikipedia. 2017. Web search engine. (October 2017). https://en.wikipedia.org/wiki/Web_search_engine

5 ISSUES

DONE:

Well done! But be aware of the warnings of the bib!

List of Figures

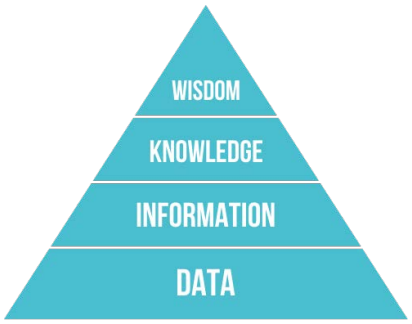

Figure 1: DIKW hierarchical model [8].

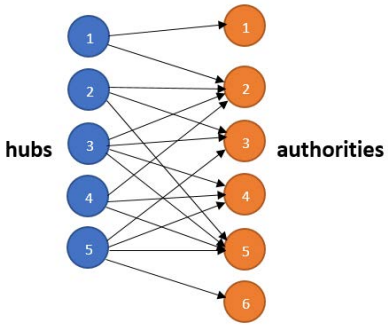

Figure 2: Hubs structure for web pages.

List of Tables

Year	Events
1990	First engine “Archie” appeared.
1994	Original Yahoo was human created catalog.
1995-2000	The classic information retrieval techniques adapted to HTML.
1998	Google founded with its link structure by using the PageRank algorithm.
2000-2005	Add context, spell check, suggestions, multiple sources.
2005-	Add optimization of complete results, topic analysis of documents, social search.

Table 1: Search engines development.