HisTrace: Building a Search Engine of Historical Events*

Lian'en Huang
Institute of Network Computing and Information Systems
Peking University
Beijing, China P.R.
hle@net.pku.edu.cn

Jonathan J. H. Zhu
Dept of Media & Communication
City University of Hong Kong
Kowloon, Hong Kong
852-27887186
j.zhu@cityu.edu.hk

Xiaoming Li
Institute of Network Computing and Information Systems
Peking University
Beijing, China P.R.
lxm@pku.edu.cn

ABSTRACT

In this paper, we describe an experimental search engine on our Chinese web archive since 2001. The original data set contains nearly 3 billion Chinese web pages crawled from past 5 years. From the collection, 430 million "article-like" pages are selected and then partitioned into 68 million sets of similar pages. The titles and publication dates are determined for the pages. An index is built. When searching, the system returns related pages in a chronological order. This way, if a user is interested in news reports or commentaries for certain previously happened event, he/she will be able to find a quite rich set of highly related pages in a convenient way.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *Information filtering, Search process*

General Terms

Design, Experimentation

Keywords

Web archive, Text mining, Replica detection

1. INTRODUCTION

The World Wide Web has been born for nearly 20 years. During the years, innumerous web pages have appeared and disappeared. Although today's search engines are very successful in helping find currently existing web pages, there still seems to be no way to search for the disappeared ones from the current Web, for example, those collected and preserved by Internet Archive [1].

Intuitively users may expect to have a search engine on collected historical web pages to work similarly as Google, indexing all the visible texts of every web page. However, this is not cost-effective. For one thing, building such a search engine will be very costly, since the number of ever existing web pages is much greater than that of currently existing web pages. For the other thing, when a web page disappeared from current Web, it often means that page is out of date, and few people probably will care about it.

Undoubtedly there are lots of valuable materials on collected historical web pages worth people searching for and mining on. The question is what kind of tools should be built for the purpose. To this end, we are building an experimental search engine, called HisTrace, which is designed for tracing historical events. To minimize the cost, a major bulk of our system is to focus only on web pages that contain formally-written articles and to provide search services only for historical events that are important and likely to be cared by many people.

Our system is based on Web InfoMall [2], a Chinese web archive developed at Peking University since 2001. As of today, it has accumulated about 3 billion Chinese web pages since early 2002, which is still growing at the rate of 1 to 2 million pages a day.

2. THE FRAMEWORK OF HISTRACE

Historical events are those important things that happened in the past. Obviously these things were broadly reported or described on historical web pages. Technically we define a historical event as an object described in some web pages, the descriptions of which likely evolve with time. It could be a war (the evolution of the war) or a person (the living history of the person).

Different from ordinary search activities, search for historical web pages introduces a new dimension: time. When a user searches for a historical event, he would probably be interested in viewing web pages in a time sequence according to how the event evolved. Our system aims to meet such needs so that when users request a query, they will get a list of returned web pages that are ordered chronologically, together with their hotness degrees.

Conceptually, our system is composed of five parts of work, described as follows.

1. **Representation of historical events (*R*)**. That is, a method to let users to describe historical events. Obviously the simplest way is to specify some keywords, just as in Google. It seems natural but difficult to characterize historical events properly in most cases. So we try a new approach to representation of historical events. In particular, we define a historical event as something that can be described in certain keywords, certain attributes, and a domain. Attributes may include "when", "where", "who" and others while a domain is composed of some descriptive words and various domain-specific high-frequency words.
2. **Selection of valuable web pages (*S*)**. The web pages we have collected contain all kinds of types, some of which are meaningless for our purpose and many others are simply spam content. A selection is therefore necessary to include those web pages that are most valuable and suitable for our

WWW 2008, April 21--25, 2008, Beijing, China.
ACM 978-1-60558-085-2/08/04.

* Supported by NSFC (60573166), HKSAR CERG (CityU 1456/06H) and City University of Hong Kong SRG (7001882), and Guangdong Key Lab Grant (CCNL200601).

system. This is an important consideration for building a low-cost system because the data volume becomes smaller. Here we select a special type of web pages as our target, called article-type web pages, of which each page should contain a formally-written article that will be discussed later. Moreover, there are also some other work needed such as spam detection and page ranking for determining web pages' importance.

3. **Match between historical events and web pages (*M*)**. After *R* is determined and *S* selected, a mapping from elements R to subsets of S is needed. Since the match should be done in real time, we need to build indexes of *S* in advance, just as in ordinary search engines. However, there is some difference for our system. For example, we don't intent to index the full text of web pages to match keywords of historical events. Instead, we index only the titles and key sentences extracted from full text, which makes the data volume smaller. By this approach, we certainly will miss some matches but we believe it helps to improve the accuracy and the focus of main themes.

4. **Determination of web pages' publication-time**. As mentioned before, our system is capable of searching on time dimension. To accomplish that objective, we need to determine the publication-time (or birth-time) of each web page. Furthermore, if one page has several replicas, what we want is the publication-time of its first copy. Web pages' crawl-time and last-modified-time (by HTTP protocol) are useful but not precise enough for the purpose. In practice, we use content extraction, link analysis and replica detection to help determine the publication-time of web pages more precisely.

5. **Display of search results**. On the Web important articles are often replicated across many web sites. As a result if a user submits a query and then gets lots of similar web pages returned, he or she will become bored. It is important to cluster the similar web pages together and select only one as a delegate to return to the user.

3. SYSTEM IMPLEMENTATION

3.1 HisTrace Architecture

The architecture of our system is shown in Figure 1. The base of our system is Web InfoMall (http://www.infomall.cn), a large-scale Chinese web archive with nearly 3 billion web pages collected. We also have built a large-scale web graph for it, based on which a ranking system has been set up.

However, we don't build our search engine directly on the whole web archive. As we focus only on article-type web pages, an article set is thus set up by extracting articles from the web archive. Replica detection is further done on the article set to remove duplicated articles, and the publication-time of each article is also computed. Compared to the original web archive, the article set is much smaller in data volume. Accordingly, a link graph of articles and its ranking system are set up.

On the upper level, a mining subsystem on articles is built. We extract some useful information from the articles, e.g., key sentences. Above all, titles, key sentences of articles and other extracted information are indexed for providing search service.

In the following two subsections, we briefly describe some key implementations of building the article set.

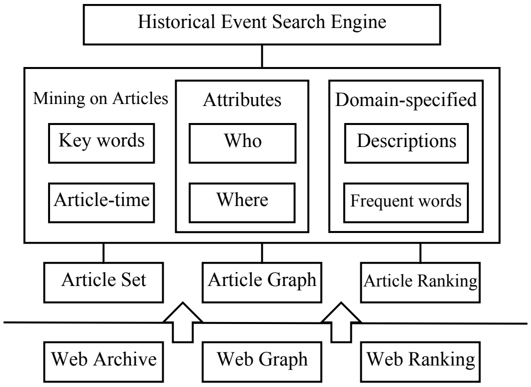

Figure 1. Architecture of HisTrace

3.2 Extraction of Articles

Articles are extracted from article-type web pages. An article-type page is a kind of topical page that contains a formal article (i.e., it has a title and sufficient amount of content). The main task here is to determine whether a web page is an article-type page. To do so, we first check HTML tag <title> to get title of the web page (or it can be got from anchortext). Then we seek for the title in the page content. If it is found, we then check to see if it follows by a sequence of continuous text. Some tests are then performed to verify if there is good content of an article.

3.3 Replica detection

An important requirement of our system is that replica detection should be precise enough; otherwise it may cause bad results returned to users. To meet the requirement, we have developed a new approach of near-duplicate detection that achieves both high precision and high recall.

Technically, our approach is based on LCS (longest common subsequence) [3] for primary similarity measurement, together with a carefully designed framework of procedures, which includes three steps: (1) clustering the web pages into sets of perhaps-similar pages to constrain the computation of LCS to occur only between perhaps-similar pages, and (2) making sketches of web pages by a filtering method to further reduce their differences before computing LCS, and (3) computing LCS on sketches and choosing its trustable portion to compute similarity.

4. CURRENT PROGRESS

Currently we have already built a preliminary system (see http://hist.infomall.cn). It runs on a data set of 68 million articles which come from Web InfoMall's 3 billion web pages. The web pages are crawled during the last 5 years, from which we have first extracted 430 million articles and a process of replica detection has been done to form the set of 68 million articles.

As for matching between historical events and articles, the current system simply matches key words of historical events with titles of the articles. Other functionalities will be added in the future.

5. REFERENCES

[1] Internet Archive, http://www.archive.org

[2] Web InfoMall, http://www.infomall.cn

[3] E. W. Myers. An O(ND) difference algorithm and its variations. Algorithmica, 1(0178-4617):251-266, 1986