§27. Development of Search Engine to Chorological Analysis of Word Appearance Frequency

Emoto, M. (NIFS), Teramachi, Y. (Polytechnic Univ.), Yamaguchi, S. (Chubu Univ.)

To estimate exact supply and demand of energy, exact prospect is necessary. However, it is difficult to prospect because they depend on various social events. Therefore, it is important to study how much one event affects the other events. In order to study the influence of each event, the authors have developed the database system to show the time history of frequency of the words appeared in the news articles.

The figure 1 shows the overview of this system. The database server collects the articles from the news site in the Internet at midnight. The collected articles are the web pages of the sites down to 4 levels from the top (the pages that can be accessed from the top pages within 4 links). The total size is about 1 million pages or 2GB per day, and it is hard to look for look for the words from the entire pages. Therefore, the authors adopted NAMAZU [1], an open source full text search engine to do this task. It registers all the words appeared in the articles into the index database. It also registers the frequency and the score of these words for each page. The score is a weighted frequency; it becomes large value when it is used in important contexts such as titles, list items, and boldfaced sentences.

The authors provide a Java applet based application to use this system interactively (Figure. 2). When the user enters the words and the period, the application draws the history graph of the frequency and scores of the words during the period. The actual search is done by the server; when the server receives the request, it looks for the index to calculate its frequency and scores. To improve the performance, the system consists of several index servers to make a cluster. When the database server receives the requests, it asks the index servers to look for the word at the same time it looks for the word by itself. The index server has only index of the data and don't have news data. Currently, the number of the index servers is three, but it is easier to increase performance by increasing the number of index servers.

Using NAMAZU, the system works fine, but the author is currently looking for alternative search engine. One of the reasons to look for another engine is time to make indexes. The index is made by a batch job from the news data. It takes about three hours, and it becomes longer as the number of the articles increases. The second reason is that the searching time becomes large when the period is longer. Therefore, a faster search engine is necessary for the future demand. One of the candidates is Hyper Estraier [2]. The indexer of Hyper Estaier is written C++ while that of NAMAZU is written by perl, so it can make an index for 20 minutes, 6 times faster than NAMAZU. Furthermore, because it uses simpler method to look for the word, it is faster than NAMAZU. However, the index size of Hyper Estaier is 1.5 GB per day while NAMAZU's index size is 200MB. Because the authors don't have enough disk storage to store all the indexes now, we haven't replaced the search engine yet.

The scoring plays an important role for full text search. The scoring used in this system is that of NAMAZU as it is. However, NAMAZU is a generic tool to find a file, and its algorithm is not necessarily suitable for this system to analyze the word frequency. For the further study, we are planning to develop a scoring algorithm for this system.

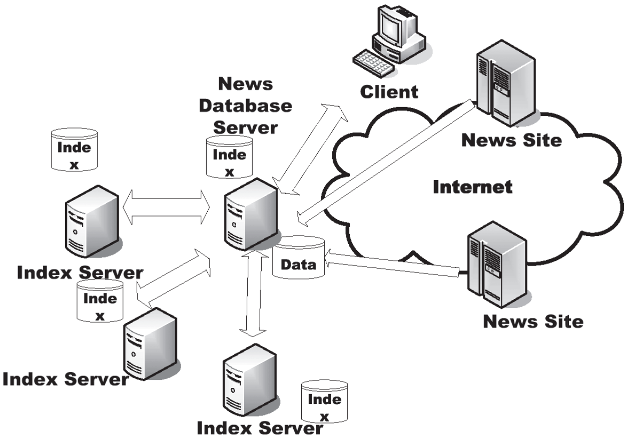
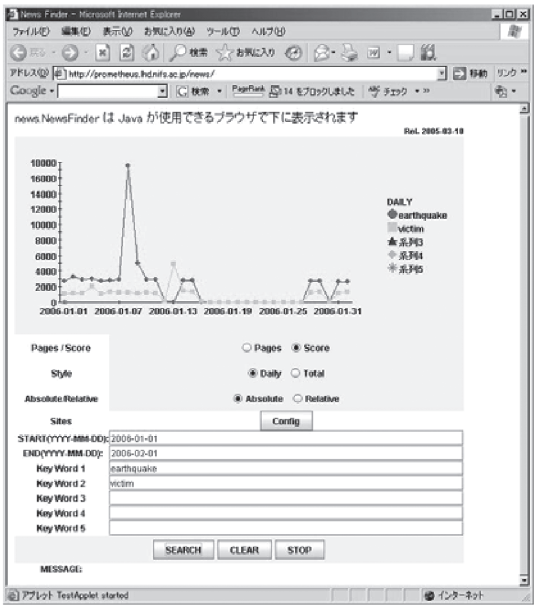


Fig1. System Overview


News Finder - Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
検索 お気に入り
アドレス(D) http://prometheus.lhd.nifs.ac.jp/news/
移動 リンク
Google 検索 PageRank 14 をブロックしました チェック
Rel. 2005-03-10
10000
16000
14000
12000
10000
8000
6000
4000
2000
0
2006-01-01 2006-01-07 2006-01-13 2006-01-19 2006-01-25 2006-01-31
DAILY
● earthquake
victim
▲ 系列3
◆ 系列4
✳ 系列5
SEARCH
CLEAR
STOP
アプレット TestApplet started
インターネット

news.NewsFinder は Java が使用できるブラウザで下に表示されます

Pages / Score ○ Pages ● Score

Style ● Daily ○ Total

Absolute/Relative ● Absolute ○ Relative

Sites Config

START(YYYY-MM-DD): 2006-01-01

END(YYYY-MM-DD): 2006-02-01

Key Word 1 earthquake

Key Word 2 victim

Key Word 3

Key Word 4

Key Word 5

MESSAGE:

Fig 2. Client Application

Reference
1) http://www.namazu.org/
2) http://hyperestraier.sourceforge.net/