OASIS at NTCIR-5: WEB Navigational Retrieval Subtask

Vitaly KLYUEV
The University of Aizu, Japan
vkluev@u-aizu.ac.jp

Abstract

We experienced negative results participating in this Subtask: the OASIS system, which is a distributed search system based on VSM and full text indexing, failed to retrieve relevant documents from the huge data set of Japanese Web pages when the number of relevant documents in the collection was relatively small.

1 Methods Used

We applied several techniques to enhance the vector space model implemented. Some of them are listed below:

1. Weights of the terms gathered around URLs were increased in a 'wave' style manner.

$$w^w_{i,u} = w_{i,u} + \frac{1}{2 * |i|} * w_{i,u} for i = \pm 1, \ldots, \pm 5,$$

   where i as a distance in words from the URL u; $w_{i,u}$ is a weight of the term.

2. A two-stage retrieval technique was used to get and merge results obtained from the servers.

3. Alternatively, the Borda-fuse algorithm was adapted. The top ten documents from each returned list were given a new score:

$$S(D^k_i) = \frac{101 - i}{N_k}, i = 1, \ldots, 10$$

   where D^k_i is *i*th ranked document retrieved by server $\hat{k}$; N_k is a number of documents retrieved by server $\hat{k}$.

2 Search Results

Only mandatory queries were submitted to the system. A retrieval process was carried out in a fully automatic way. Table 2 gives some examples.

Table 1. Parameters of the Servers

Number	Processor	Memory	Hard Drive
1	Dual Intel Xeon 2.8GHz	2GB	500GB
2	Intel 1.7GHz	1GB	1TB
3	Intel 1.8GHz	2GB	300GB

Table 2. Retrieval Statistics: Some Examples

Query number	Number of retrieved documents	Number of relevant documents in the collection
1003	5	2
1005	22	1
1008	3	5
1394	21	6
1395	40	21
1398	38	1

3 Failure Analysis

We failed to retrieve relevant documents applying an improved VSM approach to the huge dataset when the number of relevant documents in the collection was relatively small. We see the following reasons for this:

- Some bugs remain in our software.
- Probably, adding heuristics and incorporating negligible improvements to the commonly used models (VSM, probabilistic, etc.) cannot be an efficient solution to discover a small amount relevant documents and retrieve them successfully from very huge data collections. We think they cannot work well for this task.
- Scientists working in the text information retrieval area need a new language model applicable to natural languages (Japanese, English, etc.) to design new systems which can be tools for the Navi-2 Subtask.