Abstract

The amount of information available on the Web has increased dramatically. This information comes in a wide variety of media types. Therefore, to retrieve information on a topic using conventional search engines, users must search many sites from several different aspects. Thus, it is important to integrate and organize this information from the different sites. We propose an integration and organization system based on a query relaxation approach for cross-media meta-search engines. Frequently the parameters for information retrieval are too specific or exacting to generate any relevant sites. But by gradually relaxing the search terms used for information retrieval, we can solve this problem while narrowing the search to sites that are most relevant to the subject being researched. We show several examples of how the relaxation approach works as well as ways that it can be applied. We also demonstrate the advantages of our approach and future work for this research.

1. Introduction

As shown by advances in Internet technology, the number of available Web pages is increasing dramatically. Furthermore, both of the number and the type of multimedia content available on the Web have also been increasing because of the increased use of broadband services and devices for creating multimedia content (e.g., digital cameras, etc). This leads to the increase in the importance of search engines to users searching for specific content.

In the case of image contents, several search engines specifically for images (such as Google) have been developed and used. These engines have also been used as multiple search engines that can concurrently retrieve and integrate the results effectively. Although the meta-search engines have been effective for searching pertinent Web pages, they still have the following problems with image searches:

- Conventional image search engines still exhibit low precision and recall, regardless of whether the search is keyword-based (Google image search) or content-based (QBIC).
- There are not enough meta-search engines for image searches.

In this paper, we describe a query relaxation approach designed to realize a cross-media meta-search engine. The cross-media meta-search engine intuitively collects several types of information from the Web based on user-input keyword queries. The differences between conventional Web meta-search engines and cross-media meta-search engines are shown in Figs.1 and 2 are summarized as follows:

- Conventional Web meta-search engines send the user-input keyword query Q (possibly with minor changes) directly to several search engines. Modifications are not made to a user-input keyword queries. To retrieve the results conventional meta-search engines return a list of pertinent URLs with duplicates removed and with ranking scores.
- The cross-media meta-search engine is designed to collect not only Web pages, but also mixed types of multimedia contents (images, sounds etc.) by sending, query Q to several search engines, each of which is dedicated to a specific type of media content. On sending query Q, the search engine may modify and/or relax the term into Q1, Q2, ..., Qn according to the characteristics of each media type. The output of the cross-media meta-search engine is not simply a list of URLs, but a mixture of texts, images, and sounds, edited like an encyclopedia.
Naver is a type of meta-search engine that simultaneously searches for HTML pages, CG animations, images, and sounds etc., pertinent to the user-input keyword(s) and then merges the retrieved content. These results are only a collection of the retrieved information for each search engine. Naver is a cross-media meta-search engine that can retrieve information from several different media-sources from a user-input keyword query.

Cyclone is also a meta-search engine for searching ordinary Web pages. From a user-input keyword query, it searches several search engines and assembles the results into an encyclopedia-like form. Neither query modification nor query relaxation are used by Naver or Cyclone.

Our query relaxation approach intuitively transforms a user-input conjunctive keyword into sets of sub-query pairs, in which the sub-queries are automatically routed to the various types of search engines. This idea was motivated by the fact that conventional image search engines, such as Google do not look for images for conjunctive keyword queries consisting of many keywords. Evaluation of our query relaxation approach shows that it can dramatically improve the recall ratio for image retrievals. We also propose a system of assembling the various types of retrieved multimedia information into an encyclopedia-like form.
2.2 Answers by Query Relaxation

Answers for query Q are retrieved as unions of all the sub-query tuples:

\[ Ans(Q) = (Ans(k_1, \ldots, k_n, E_2)) \]
\[ \cup (Ans(k_1, E_1) \cap Ans(k_2, \ldots, k_n, E_2)) \]
\[ \cup (Ans(k_1, E_1) \cap Ans(k_2, \ldots, k_n, E_2)) \]
\[ \cup \ldots \]
\[ \cup (Ans(k_1, k_2, E_1) \cap Ans(k_2, \ldots, k_n, E_2)) \]
\[ \cup (Ans(k_1, k_2, E_1) \cap Ans(k_2, k_3, \ldots, k_n, E_2)) \]
\[ \cup \ldots \]
\[ \cup (Ans(k_1, \ldots, k_n, E_1)) \]

where Ans(Q, E_i) means the answers of the query Q by the i-th search engine. In this case, E_1 is a text search engine and E_2 is an image search engine. The engines return URLs which match the queries.

To compute \( (Ans(k_1, E_1) \cap Ans(k_2, \ldots, k_n, E_2)) \), \( Ans(k_1, E_1) \) and \( Ans(k_2, \ldots, k_n, E_2) \) are processed separately. Then the intersection of the answers is calculated. The answers are retrieved in the same way for all of the tuples. Finally, Ans(Q) can be retrieved as unions of all the answers.

We would now like to demonstrate the results obtained using the query relaxation approach with the following query:

\[ Q = q_1 \land q_2, \text{ where } q_1 \text{ is “Mt. Fuji” and } q_2 \text{ is “snow”.} \]

Fig. 3 shows venn diagrams of the results of query Q. The left diagram shows the results from a conventional meta-search engine. The right diagram shows the results using the query relaxation approach. In the left diagram, the hatched areas show the results for \( Ans(q_1, q_2, E_{text}) \), and the right hatched areas show the results for \( Ans(q_1, E_{text}) \land Ans(q_2, E_{img}) \), where \( E_{text} \) is the text search engine and \( E_{img} \) is an image search engine. The top hatched areas on the right diagram shows the results for \( (Ans(q_1, E_{text}) \land Ans(q_2, E_{img})) \). The bottom hatched areas shows the results for \( (Ans(q_2, E_{text}) \land Ans(q_1, E_{img})) \). Therefore, using proposed query relaxation, the possibility for obtaining pertinent results is increased.

3. Prototype System

3.1 Query Relaxation Experiment

We performed a simple experiment to relax the query. The number of keywords is set at three. We randomly chose keywords from the headlines in the news Web pages. We define the degree of relaxation as the number of keywords in a user-input query Q that are actually used by a Web text search engine. That is, when three of the keywords are used for an image search engine, the degree of relaxation is considered to be zero. When two keywords are used for an image search engine and the third keyword is used for a text search engine, the degree of relaxation is considered to be one. We compare the results of 0 of relaxation, to 1 of relaxation, and 2 of relaxation respectively.

3.2 Experimental Results

The results of the experiment are shown in Fig.4. The recall and precision graphs for each keyword are also shown in Fig.4. The precision and recall are displayed for every degree of relaxation. Precision is defined as the average precision for every degree of relaxation. Recall is defined as the total number of pertinent pages found for each of experiment. It is assumed that the recall for 2 of relaxation is 100%. Therefore, the results are relative.

The recall ratio for the results increases from 0 to 1 for every degree of relaxation. That is, increasing the number of keywords for the text search engine is an effective means of obtaining pertinent Web pages. The recall ratio from 1 to 2 did not increase significantly.
In the graph, there are two patterns. Pattern (a), (b), and (c), are normal: the recall ratio is increased and the precision ratio is decreased according to the degree of relaxation. On the other hand, pattern (d) and (e) show from 0 to 1 of relaxation with and increased precision and recall ratio that is particularly apparent in pattern (e). No answers were obtained at 0 of relaxation, but by relaxing the degree, many answers were obtained. This pattern is usually observed in queries where the number of answers falls below zero degrees.

3.3 Answer Integration and User interface

Figure 5 shows a user interface for our prototype system. Retrieved images and text are displayed in a window in combinations of related images and texts. That is, this system retrieves not only pertinent Web pages, but it can also extract pertinent portions of Web pages. The weighting term ‘tf-idf’ is used to extract text from the Web pages. Pertinent paragraphs are selected by weight and information is then extracted and integrated with the related images.

4. Conclusion and Future Work

We have described and demonstrated a cross-media meta-search method that uses a query relaxation approach to obtain pertinent Web pages. We have also demonstrated our approach and shown examples of how to calculate it. We believe that the query relaxation approach is an efficient method for image searches. While this prototype system is currently being used for image searches, we are working on expanding its application to include various cross-media types.

5. References


