Karl Aberer Zhiyong Peng
Elke A. Rundensteiner Yanchun Zhang
Xuhui Li (Eds.)

LNCS 4255

Web Information Systems – WISE 2006

7th International Conference on
Web Information Systems Engineering
Wuhan, China, October 2006, Proceedings

Springer

Karl Aberer Zhiyong Peng
Elke A. Rundensteiner Yanchun Zhang
Xuhui Li (Eds.)

Web Information Systems – WISE 2006

7th International Conference on
Web Information Systems Engineering
Wuhan, China, October 23-26, 2006
Proceedings

Springer

Lecture Notes in Computer Science 4255

Table of Contents

Keynote Papers

Session 1: Web Search

Session 2: Web Retrieval

Session 3: Web Workflows

Session 4: Web Services

Session 5: Web Mining

Session 6: Performant Web Systems

Session 7: Web Information Systems

Session 8: Web Document Analysis

Session 9: Quality, Security and Trust

Session 10: Semantic Web and Integration

Session 11: XML Query Processing

Session 12: Multimedia and User Interface

Internet-Scale Data Distribution: Some Research Problems

M. Tamer Özsu

David R. Cheriton School of Computer Science, University of Waterloo
tozsu@uwaterloo.ca

Abstract. The increasing growths of the Internet, the Web and mobile environments have had a push-pull effect. On the one hand, these developments have increased the variety and scale of distributed applications. The data management requirements of these applications are considerably different from those applications for which most of the existing distributed data management technology was developed. On the other hand, they represent significant changes in the underlying technologies on which these systems are built. There has been considerable attention to this issue in the last decade and there are important progress on a number of fronts. However, there are remaining issues that require attention. In this talk, I will review some of the developments, some areas in which there has been progress, and highlight some of the issues that still require attention.

K. Aberer et al. (Eds.): WISE 2006, LNCS 4255, p. 1, 2006.

Towards Next-Generation Search Engines and Browsers – Search Beyond Media Types and Places

Katsumi Tanaka

Graduate School of Informatics, Kyoto University
tanaka@dl.kuis.kyoto-u.ac.jp

Abstract. In this keynote talk, the author will describe concepts and technologies for next-generation search engines and browsers for searching and browsing contents beyond media types and places. Currently, digital content are represented by different media such as text, images, video etc. Also, digital content are created, stored and used on a variety of places (devices) such as independent digital archives, World Wide Web, TV HDD/DVD recorders, personal PCs, digital appliances and mobile devices. The viewing styles of these content are different. That is, WWW pages are accessed and viewed in an active manner such as a conventional Web browser (reading, scrolling and clicking interface). On the other hand, TV content are accessed and viewed in a passive manner. As for searching these "ambient multimedia contents", currently, many commercial search engines cover only WWW content and personal PC contents, called "desktop search".

First, the author describes research issues necessary for searching "ambient multimedia contents". The main research issues are (1) cross-media search, (2) ranking methods for contents without hyperlinks, and (3) integration of search results. As for cross-media search, the author describes query-free search, complementary-information retrieval, and cross-media meta-search.

Second, the author describes ways of browsing "ambient multimedia content". The main topics of the second part are new browsers by media conversion of digital content, concurrent and comparative browsers for multiple contents. For example, the proposed browsers have an ability to automatically convert Web content into TV content, and vice versa.

The last part of the talk is concerned with mining metadata owned by search engines and its usage for computing the "trustness" of the searched results.

K. Aberer et al. (Eds.): WISE 2006, LNCS 4255, p. 2, 2006.

Building a Domain Independent Platform for Collecting Domain Specific Data from the Web

Lizhu Zhou

Department of Computer Science and Technology, Tsinghua University
dcszlz@tsinghua.edu.cn

Abstract. World Wide Web has become a major information resource for both individuals and institutions. The freedom of presenting data on the Web by HTML makes the information of same domain, such as sales of book, scientific publications etc., be present on many Web sites with diverse format. Thus to collect the data for a particular domain from the Web is not a trivial task, and how to solve the problem is becoming a trendy research area. This talk first gives an overview of this new area by categorizing the information on the Web, and indicating the difficulties in collecting domain specific data from the Web. As a solution, the talk then continues to present a stepwise methodology for collecting domain specific data from the Web, and introduce its supporting system SESQ which is a domain independent tool for building topic specific search engines for applications. The talk shows full features of SESQ by two application examples. In conclusion, the talk briefs further research directions in this new Web data processing area.

K. Aberer et al. (Eds.): WISE 2006, LNCS 4255, p. 3, 2006.

A Web Search Method Based on the Temporal Relation of Query Keywords

Tomoyo Kage and Kazutoshi Sumiya

University of Hyogo
1-1-12 Shinzaike-honcho, Himeji, 670-0092, Hyogo, Japan
nd05w005@stshse.u-hyogo.ac.jp
sumiya@shse.u-hyogo.ac.jp

Abstract. As use of the Web has become more popular, searching for particular content has been refined by allowing users to enter multiple keywords as queries. However, simple combinations of multiple query keywords may not generate satisfactory search results. We therefore propose a search method which automatically combines query keywords to generate queries by extracting the relations among query keywords. This method consists of two Web search processes: one to determine the temporal relations between query keywords, and one to generate queries based on the obtained temporal relations. We discuss these two processes along with experimental results and implementation issues regarding a prototype system.

Keywords: Information Retrieval, Temporal Relation, Web Archive, Query Generation.

1 Introduction

Current Web search engines based on the relations between keywords and Web pages only consider the presence or absence of keywords. For that reason, if users input some keywords as query keywords to a Web search engine, it outputs only Web pages that include all the query keywords. For example, suppose a user inputs {*cherry blossoms*, *autumn leaves*, *trip*} to a Web search engine. Current search engines will output Web pages that include all the query keywords; i.e., pages containing information about sightseeing. However, in some cases the user is likely to want information from travel agency pages which provide tour information or from personal sites where trip topics are discussed.

In this study, the automatic generation of meaningful queries using OR bindings makes it possible to get useful Web pages. We determine the relation between query keywords on Web pages as a temporal relation to generate the queries. This temporal relation is extracted from a time series of Web pages. The time series of Web pages we use are collected from Web archives[1] .

[1] http: //www.archive.org/

K. Aberer et al. (Eds.): WISE 2006, LNCS 4255, pp. 4–15, 2006.

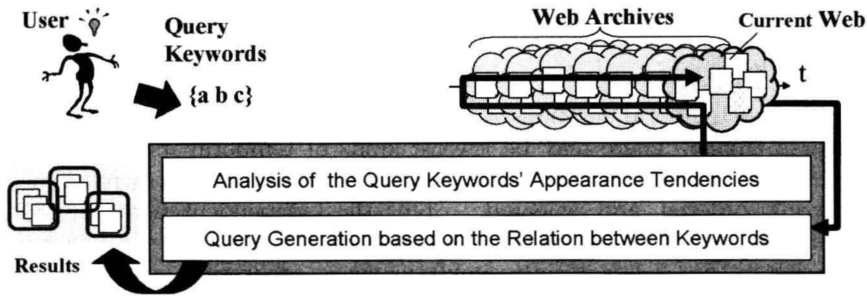

Fig. 1. Concept of Our Approach

In this paper, we propose using the temporal relations between query keywords to generate query keyword combinations (Fig. 1). We analyze the appearance tendencies of query keywords, generate queries by extracting the temporal relation, and then get Web pages which cannot be obtained by current engines.

Where the query keywords are {*cherry blossoms*, *autumn leaves*, *trip*}, the characteristic of {*cherry blossoms*, *autumn leaves*} alternately appearing in a time series of Web pages is extracted, and queries returning {R(*cherry blossoms* $\wedge$ *trip*) $\vee$ R(*autumn leaves* $\wedge$ *trip*)} are generated. That is, the system will find that there is a fairly small chance of these keywords appearing at the same time. Users can get travel agency pages including tour information through the query. Furthermore, this method is much more efficient than a method that generates all combinations joined by AND binding or OR binding. This is because only meaningful queries are generated by using the relation between query keywords.

Related work on query generation includes the multimedia retrieval of Web pages by query relaxation [8] and interval glue operations for video data retrieval [9]; these methods differ from ours, however, because their retrieval target is different. Regarding studies using time series pages, approaches specialized for the purpose of summarization [1][2][3][7] have been proposed. Furthermore, [5][6] aim at search engine ranking. They deal with temporal Web pages with the same target as our method, but the goals and methods differ from ours. Research on the temporal relation of keywords includes [10] and [4]. Temporal characteristics are extracted from the query logs in the search engine; however, we extract these from Web logs, and so their approach differs from ours. A search engine which applies clustering to search results is also available [11]. This search engine classifies subtopics for query keywords by analyzing Web page summaries, while we use the temporal features of query keywords.

The remainder of this paper is organized as follows. Section 2 describes the process to determine the temporal relation that we propose, and Section 3 describes the query generation. Section 4 looks at an experiment to test the proposed method and the design of the prototype system. In Section 5, we summarize this work.

2 Temporal Relation of Query Keywords

2.1 Definition

We define the temporal relations of query keywords that users input into a Web search engine. There may be several kinds of temporal relation between the keywords. We make connections between the relations. Three relations — co-occurring, ordered, and repeated — plus the independent relation, are all basic temporal relations. The strict co-occurring relation is a special case of co-occurring in that it is a strictly simultaneous co-occurrence, while the cyclic relation is a special case of repeated in that the time intervals are exactly the same. The temporal relations are defined as follows.

Co-occurring. The co-occurring relation is where keywords a and b both appear at the same time on time series pages. General co-occurrence means that multiple keywords appear on a Web page. However, we do not pinpoint a page because the scale is time. Consequently, if keywords a and b appear at more or less the same time, we say that they have a co-occurring relation. For example, a keyword pair {*SARS*, *corona*} has a co-occurring relation. In this case, a tight relation exists between the keywords, and a strict co-occurring relation is classified as a strong relation of this type. This is described in more detail below.

Ordered. The ordered relation is where keywords a and b are order dependent with respect to each other; i.e., one of them appears before the other in time series pages. That is, with two keywords, either "a appears before b" or "b appears before a". A concrete example of this relation is that of the keyword pair {*SBC*, *AT&T*}. In this way, a "causal" semantic relation is included and the relation expresses a topic change. We classify ordered as repeated, though, when there is a strong relation, as described in more detail below.

Repeated. The repeated relation is where keywords a and b appear in a circular pattern in chronological order; i.e., a, and b alternately appear on time series pages. For instance, the keywords pair {*disaster*, *volunteer*} or {*remodeling*, *reopening*} has this relation. A repeated relation includes the ordered relation because a keyword pair of the repeated relation has the ordered relation and appears two or more times. A cyclic relation is classified as a strong repeated relation. This is described in more detail below.

Independent. The independent relation is where keywords a and b appear independently on time series pages. That is, they appear in random order with respect to each other; for example, such a keyword pair might be {*pumpkin pie*, *recipes*}. We define the independent relation as a residual which includes relations other than the three relation types described above.

Strict Co-occurring. The strict co-occurring relation is where keywords a and b both appear on the same page from among the time series pages, and it is strongly related to the co-occurring relation described above. For example, {*Halloween*, *costume*} has such a relation.

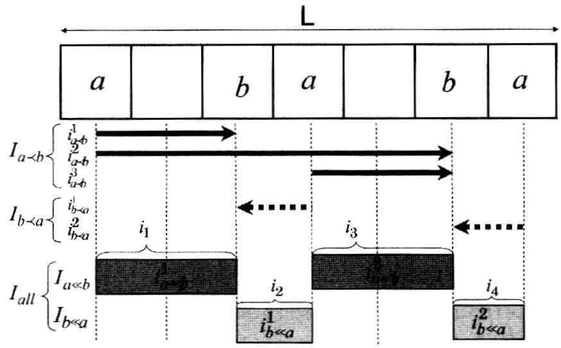

Fig. 2. Extraction of Intervals

Cyclic. The cyclic relation is where keywords a and b appear temporally and periodically, and it is strongly related to the repeated relation described above. This relation differs from the repeated one in that the keywords appear regularly. In particular, the time length of a to b is constant, as it is for b to a. That is, the time lengths of intervals a to b and b to a are approximate. For example, in the case of {*cherry blossom, autumn leaves*}, the time interval of spring to autumn is the same every year, as it is for autumn to spring. The same would hold for events (e.g.,{*Halloween, Christmas*}) which usually appear periodically.

2.2 Determination

In this study, we determine the temporal relation between the keyword pairs on each URL. We generate keyword pairs and operating intervals around a keyword pair. We then identify each pair as having either a co-occurring, ordered, or repeated relation. The determination is done in the order of co-occurring, ordered, and cyclic. The method of determination for a keyword pair a, b is described below.

First, we extract intervals $I_{a \prec b}$ and $I_{b \prec a}$ to determine the temporal relation between query keywords. $I_{a \prec b}$ is a set of $i_{a \prec b}$, and $i_{a \prec b}$ is an interval from "a page that includes a" to "a page that includes b". Correspondingly, $I_{b \prec a}$ is a set of $i_{b \prec a}$, and $i_{b \prec a}$ is an interval from "a page that includes b" to "a page that includes a". We then extract intervals $I_{a \ll b}$ and $I_{b \ll a}$. $I_{a \ll b}$ is a set of $i_{a \ll b}$, and $i_{a \ll b}$ is an interval satisfying a condition that a appears earlier than b. As before, $I_{b \ll a}$ is a set of $i_{b \ll a}$, and $i_{b \ll a}$ satisfies a condition that b appears earlier than a. The extraction process of $I_{a \ll b}$ is as follows (Fig. 2).

- **STEP 1** Extract $I_{a \prec b}$
- **STEP 2** Remove $i_{a \prec b}$ which can be regarded as having $i_{b \prec a}$

STEP 2 is necessary because $I_{a \ll b}$ is not extracted from the parts of intervals where $i_{b \prec a}$ are of $i_{a \prec b}$.

In addition, we determine the temporal relations through extracted intervals. The definition of Fig. 2 and the algorithm for determining the relations are explained in the following. We define the intervals of time-series pages as follows. L is the time length of time-series pages. $I_{a \prec b}$, $I_{b \prec a}$, $i_{a \prec b}$, $i_{b \prec a}$, $I_{a \ll b}$, $I_{b \ll a}$, $i_{a \ll b}$, and $i_{b \ll a}$ are as having been defined. The set of $I_{a \ll b}$ and $I_{b \ll a}$ is I_{all}, and $I_{all} = \{i_1, i_2, ..., i_n\}$.

//Determination of Co-occurring Relation
if $\left(\frac{cnt(less\_time(I_{a \prec b}))+cnt(less\_time(I_{b \prec a}))}{cnt(I_{a \prec b})+cnt(I_{b \prec a})} \geq \alpha\right)$
  then **//Determination of Strict Co-occurring Relation**
    if $\left(\frac{cnt(time\_0(i_{a \prec b})}{cnt(I_{a \prec b})+cnt(I_{b \prec a})} \geq \alpha\right)$
      then
        strict co-occurring
      else
        co-occurring
    fi
  else goto Determination of Ordered Relation
fi

Fig. 3. Determination of Co-occurring Relation

Co-occurring Relation. If the total of co-occurring keywords among all keywords that appear is large, we determine that the relation is co-occurring. We set a threshold value which can be regarded as the same time, and if the total within which both keywords appear is more than the threshold value, the relation is co-occurring. The algorithm is shown below (Fig. 3).

We define the function as follows. $time(i)$ returns the time length of i. $cnt(I)$ returns the total of intervals included in I. $less\_time(I)$ returns intervals (included in I) whose time length is less than a threshold value. $time\_0(i)$ returns the total of i whose time length is zero. α is a threshold value.

We use $i_{a<b}$ and $i_{b<a}$, which can be regarded as keyword-appearing intervals, to determine co-occurring. The appearing intervals of keyword pair a and b are calculated by $time(i_{a<b})$, and the total of co-occurring keywords is a summation of $i_{a<b}$ and $i_{b<a}$ where the value is less than the threshold. We extract the ratio of the total of co-occurring to the total of appearing, and a relation is regarded as co-occurring if the value is larger than threshold α. When a relation is found to be co-occurring, we determine whether it is strict co-occurring[2]. If the ratio of total of $time(i_{a<b}) = 0$ to the total of appearing is larger than threshold α, the relation is strict co-occurring.

Ordered Relation. If the relation is not determined as co-occurring, ordered relation determination is then done. If the order of a keyword pair appearing on time series pages is unique, the relation is determined as ordered. The method to determine an ordered relation is shown in Fig. 4 and proceeds as follows.

[2] We consider this to mean that both keywords appear on the same pages.