使用 Fuzzy Concept Map 之超文件資訊撷取
Hypertext Information Retrieval using Fuzzy Concept Map
林典論 張明治 賀嘉生
中原大學資訊工程學系
中華民國
e-mail : lde1976@ms6.hinet.net
maiga@ms2.hinet.net
jsheh@cycs01.ice.cycu.edu.tw

ABSTRACT
As computer network develops rapidly and
becomes more and more popular, plenty of
information appears over Internet and grows
unimaginably. Especially with the vast number
of information resources available on World
Wide Web today, it is time-consuming for users
to seek out desired information among so many
HTML pages.
This paper proposes a Hypertext
Information Retrieval Agent (HIRA), which
uses Fuzzy Concept Map and modified
Dijkstra's shortest path algorithm to achieve the
task of making searching information on World
Wide Web more quickly and efficiently.
Through constructs a fuzzy concept map to
specific knowledge domain and uses modified
Dijkstra's algorithm to calculate the related
fuzziness on fuzzy concept map according to
max-min principle, related knowledge can be
retrieved and internalized by learners more
easily.

Keywords : Information Retrieval, Concept
Map, World Wide Web, Hypertext, Fuzziness

Introduction
With the vast number of information
resources available on World Wide Web today,
it is time-consuming for users to seek out desired
information among so many HTML pages. Some
papers offer solutions to the problem by building
information retrieval agents that help users to
find out information efficiently on specific
domains.

Several of these information retrieval
agents solve the problem by constructing
concept maps on specific domains. A
concept map of a specified domain is a map that
describes relationships between plenty of
concepts in the domain. As a result, if a user is
searching for information about a concept, then
the information retrieval agent will find out the
result according to the specified concept map. In
this manner, users can rapidly get the
information they need.
The combination of information retrieval agent and a specific concept map indeed helps a lot. But adding fuzziness to information retrieval agent will let the agent more powerful on information retrieval. In this paper, fuzzy concept map is proposed to provide a more efficient solution.

In the remaining sections, Section 2 gives a detailed description of hypertext information retrieval problem. Section 3 provides the analysis of related fuzziness and accumulated information. Section 4 introduces the modified Dijkstra's shortest path algorithm, the calculation of accumulated information, and the overall algorithm. In Section 5, the implementation of information retrieval agent with fuzzy concept map is presented. In Section 6, the summary of this paper and future work are given.

Hypertext Information Retrieval Problem

As computer network develops rapidly and becomes more and more popular, plenty of information appears over the Internet and grows unimaginably. Especially with the fast development of World Wide Web, many educational institutes, colleges, companies, and associations offer information about themselves in the form of hypertext that contains texts, images, animation, sound, and movies. Such kinds of homepages are very vivid and can cause user's active reading. As a result, readers may learn many things through network.

HIR (Hypertext Information Retrieval), the retrieval of information among large number of homepages, provides an efficient approach for users to obtain the information they want over large collections of hypermedia material. HIR is becoming more and more important in modern time because it provides the combined searching and browsing paradigm of information discovery.

Information filled in the cyberspace of World Wide Web is various and colorful. Through simply clicking the hyperlink in a hypertext, we can easily obtain any kind of information in this cyberspace. Based on such aspect of hypermedia information, the relationships among information can be used to constitute a directed graph of an information space in WWW.

- $S_{ir}(U, H)$: directed graph of information cyberspace, denoting information space.
- $U_i$: node set of information pieces (URLs) on the WWW, could be any kind of media.
- $H_i$: directed edge pointing from $U_i$ to $U_j$.

The hyperlinking relationships among hypermedia information make an information architecture, as Figure 1 shown.

![Figure 1: Information architecture of WWW](image)

In general, one hypertext form of a homepage on WWW will contain several key-concepts. At the same time, there are many concepts in a specified problem domain. These concepts will interlink other concept and a concept map then is constructed necessarily for describing those relationships among concepts. The relationship between two concepts can be figured out through observing those linked edges. Because the concept map is a connected graph, even two concepts have no direct connected edge, they are still related in some sort of degree. Therefore, a fuzzy concept map shown as Figure 2. Symbols appeared in Figure 2 are defined below:

- $C_i$: node of concept, denoted by a name (string).
- $R_i$: undirected edge of the relationship between $C_i$ and $C_j$.
- $F_i$: the fuzziness of $R_i$. 
Since the hypertext of a Web page includes many concepts, the reader has to skim over the contents to comprehend the embedded meaning. This takes much time. If an agent could judge whether the contents of the page we visit are those we want, it will be very convenient for us because it saves much time for us to check up the page. This paper develops a smart agent, called HIRA (Hypertext Information Retrieval Agent), to achieve such a task. The information retrieval problem that HIRA encounters is: Given an assigned starting hypertext node, an assigned concept, and a fuzzy membership threshold, HIRA has to find out all qualified Web pages that are related to those concept with fuzzy membership greater than the assigned threshold. The following is the conditions that HIRA faces:

- **Known**: fuzzy concept graph, $C_{\text{net}}$ (C, R, F).
- **Given**: assigned concept $C_k$ starting information node $V_n$ and fuzzy membership threshold $f_n$.
- **Find**: all information (URLs) from Information Cyberspace $S_{\text{net}}$.

Offering these three parameters, HIRA will help a user look for more related articles about the concept you assigned.

Analysis of Hypertext Information Retrieval

According to the definition of fuzzy concept map in the previous section, the fuzzy relationship between two relative concepts can be found out. For those adjacent concepts, the relative fuzzy relationship can be directly calculated from $F_{ij}$. When there exists exactly a single path between two concepts, then the relative fuzziness between them is the minimal fuzziness of all edges in the path, as Figure 3 shown.

$$F_{nj} = \max_{\text{path } p \in C_{\text{net}}} \min_{\text{edge}(pq) \in Path} F_{pq}, \quad (1)$$

In Eq.(1), the relative fuzziness for each path is first computed and then the maximum of all the values in paths is chosen as the final relative fuzziness. In Figure 4, there are three possible related paths between $C_i$ to $C_j$. By using fuzzy operations, the maximal fuzziness of possible paths can be computed. Therefore, the appropriate path between two concepts is chosen by Eq.(1).

Given any hypertext $U_k$, there may be many concepts that are in the same fuzzy concept map as the assigned concept $C_k$. Therefore, each of them have relative fuzziness for $C_k$. For example, if a concept $C_j$ appears in hypertext $U_k$, then $F_{kj}$ is the relative fuzziness between $C_k$ and $C_j$. Let $\theta_j$ be the frequency of occurrence of concept $C_j$ in a hypertext $U_k$, that is, concept $C_j$ appears $n$ times in $U_k$ with $n \geq 1$, then the accumulated related information in $U_k$ can be calculated by

$$\text{related information} = \sum_{j} F_{kj} \cdot \theta_j \cdot n$$
Eq. (2) illustrates the way the accumulated related information $F_i$ in $U_i$ is calculated. Take concept $C_j$ as an example, $\theta_j$ represents its total counts appeared in hypertext $U_h$ if $C_j$ appears more frequently in $U_i$, $1 - (1 - F_i)^{\theta_j}$ will become smaller and smaller, and the corresponding information $1 - (1 - F_i)^{\theta_j}$ will be accumulated more. Then the total information $F_i$ embedded in $U_i$ will be finally selected as the maximal information of $C_j$’s. For each hypertext $U_h$, it is necessary to calculate its accumulated related information $F_i$ because $F_i$ will be considered as a selection criterion.

After $F_i$ has been calculated, it then be compared to the assigned fuzzy relationship threshold $f_{th}$ and if

$$F_i \geq f_{th},$$

(3)

the hypertext $U_i$ will be considered as highly related to $C_s$ and the hypertext $U_i$ will become qualified Web page for user’s requirement. On the contrary, if $F_i < f_{th}$, then $U_i$ will not be taken into consideration because of lacking in qualification.

**Designing Algorithm for HIRA**

Relative fuzziness between the nodes in a fuzzy concept map $F$ and one designated concept $C_s$ can be calculated by Eq. (1). It is implemented as a function $F_s = F2Ca(F,C_s)$ through modifying Dijkstra’s shortest-path algorithm, which is used to solve the single-source shortest path problem on a weighted, directed graph. In a similar way, the fuzzy information $F_i$ in an information node $U_i$ can then be calculated by Eq. (2), which is implemented as the function $F_i = I_{\text{related}}(U_i, F_s)$.

For every information node, its related information about the assigned concept is calculated and checked whether it’s a qualified node. Then, the whole algorithm adopted in HIRA is presented as follows:

(0) Given $C_{fat}(C,F), C_s, U_i, f_{th}$

(1) Let $N = 0, U_{\text{found}} = \{ \}, F_{\text{found}} = \{ \}$

(2) $U2\text{find} = \{ U_i \}$

(3) $F_s = F2Ca(F,C_s)$

(4) do for each $U_i$ in $U2\text{find}$

(4a) $U2\text{find} = U2\text{find} \setminus \{ U_i \}$

(4b) $F_i = I_{\text{related}}(U_i, F_s)$

(4c) if $F_i > f_{th}$ then

(4c1) $N = N + 1$

(4c2) $U_{\text{found}} = U_{\text{found}} \cup \{ U_i \}$

(4c3) $F_{\text{found}} = F_{\text{found}} \cup \{ F_i \}$

(4c4) search all hyperlinks $\{ U_k \}$’s in $U_i$

(4c5) $U2\text{find} = U2\text{find} \cup \{ U_k \}$

(4c0) end if

(4z) end do

Here, several variables used are defined:

- $U_{\text{found}}$: the set of qualified hypertexts.
- $F_{\text{found}}$: the set of accumulated related information for each element in $U_{\text{found}}$.
- $U2\text{find}$: the set of hypertexts to be searched.

**HIRA Implementation on CORBA**

Hypertext Information Retrieval Agent is running on WWW and is implemented as Java Applet that can be downloaded from the WWW server anywhere at any time. It’s developed using IONA’s OrbixWeb that are used to develop distributed object programming and is compiled in Java Development Kit 1.1.3.

HIRA is designed in the architecture of distributed objects, CORBA (Common Object Request Broker Architecture). As Figure 5 shows, the downloaded applet acts as CORBA client, and the IRA CORBA server is located on a remote machine. After users fill some necessary parameters, the client applet makes request to the server and waits for its response. On the other hand, when receiving requests from client, the server deals with retrieving information according to the transmitted parameters from every client, and sends the result of processing back to clients.

Figure 6 is an example of information retrieval on the Java programming language. When browsing WWW servers with HIRA, the applet will be downloaded to user’s machine and starts running. When the user fills the three parameters and presses the Start button, the
client applet makes a function call on remote server and waits for the result. Figure 7 shows the result received from the server. It contains authorized hypertext about the assigned concept. Just click any hyperlink, and relative articles will be presented on browser.

Figure 5: HIRA System Architecture

In the future, HIRA will be designed to be a practical and powerful agent that possesses the following functionality:

- Using neural network learning strategy to reconstruct fuzzy concept map in the process of information retrieval.
- Enable multiple concepts and support logic combination of concepts, e.g., C1 and C2, C1 or C2, C1 − C2.

References


Conclusion

HIRA (Hypertext Information Retrieval Agent) constructs a specific fuzzy concept map and uses modified Dijkstra’s algorithm to calculate the related fuzziness on fuzzy concept map according to max-min principle. Besides, it is implemented on WWW so that it can be conveniently used. By means of HIRA, users can quickly find the information they want without navigating many pages and spending more time reading them. Moreover, when in combination with WWW searching engine, HIRA can retrieve plenty of useful information on many domains if the agent has construct fuzzy concept maps for a lot of domains. In the near future, users navigating on WWW can get information they need at leisure because at that time HIRA has become a popular and widely used tool.

Authors

林典諭 (Den-Yuh Lin)，中原大學資訊工程學系學生。

張明治 (Maiga Chang)，中原大學資訊工程學系碩士班學生。
賀嘉生 (Jia-Sheng Heh, Ph.D.), 中原大學資訊工程學系副教授。