Collaborative Web Browsing Based on Ontology Learning from Bookmarks

Jason J. Jung, Young-Hoon Yu, and Geun-Sik Jo

Intelligent E-Commerce Systems Laboratory,
School of Computer Engineering, Inha University,
253 Yonghyun-dong, Incheon, Korea 402-751
j2jung@intelligent.pe.kr, yhyu@eslab.inha.ac.kr, gsjo@inha.ac.kr

Abstract. This paper proposes the collaborative web browsing system sharing knowledge with other users. We have specifically focused on user interests extracted from bookmarks. A simple URL based-bookmark is provided with structural information by the conceptualization of the ontology. Furthermore, ontology learning based on a hierarchical clustering method can be applied to handle dynamic changes in bookmarks. As a result of our experiments, with respect to *recall*, about 53.1% of the total time was saved during collaborative browsing for seeking the equivalent set of information, as compared with single web browsing.

1 Introduction

Recently, in order to search relevant information, navigating in this overwhelming web environment is a lonely and time-consuming task [1]. There have been many kinds of studies to handle this problem such as the personal assistant agent [2]. Collaborative web browsing is an approach whereby users share knowledge with other like-minded neighbors while searching information on the web. When communicating with the others, we can have many kinds of experiences and gain knowledge such as which searching method is more useful and which steps they needed in order to search a certain piece of information. The representative collaborative browsing systems are *Let's Browse* [5], ARIADNE [6], WebWatcher [3], and BISAgent [4].

Recognizing what a user is interested in is very important in collaborative web browsing when querying relevant information from other users and helping with searching tasks. This paper proposes the extended application of a BISAgent, which is a bookmark-sharing agent system based on a modified *TF-IDF* scheme without considering user preference. According to the GVU's survey, nowadays there is no doubt that the number of bookmarks has increased more than ever. This means that the set of bookmarks in a user's folder can be considered to be enough to infer user interests [10]. Due to the lack of semantic information from simple URL-based bookmarks, we are focusing on a way of conceptualizing them by referring to ontology. When the structural information for users' bookmarks is provided, not only the precision but also the reliability of the extraction of user preferences can be improved.

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3038, pp. 513–520, 2004.

2 Ontology Learning from Bookmarks

An ontology is a specification of a conceptualization, which plays a role in enriching semantic or structural information [7]. In addition, a bookmark means URL information about a web site that a user wants to remember and visit again during web browsing. We try to analyze not only his bookmarks but also semantic information that each bookmark implies. Thereby, ontology is applied to conceptualize the simple URL-based bookmarks, and more importantly, hierarchical clustering is exploited to learn these conceptualized bookmarks, as shown in Fig. 1.

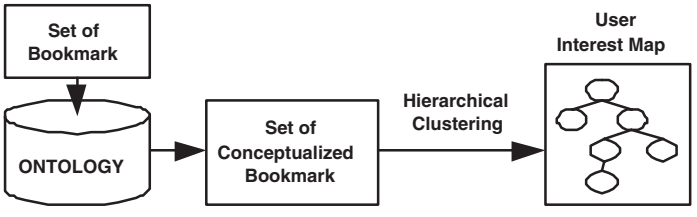

Fig. 1. Hierarchical Clustering of bookmarks

Ontology learning has four main phases which are import, extract, prune, and refine [9]. We are focusing on extracting semantic information from bookmarks based on hierarchical clustering, which is the process of organizing tree structures of objects into groups whose members are similar in some way [8]. The tree of hierarchical clusters can be produced either bottom-up, by starting with individual objects and grouping the most similar ones or top-down, whereby one starts with all the objects and divides them into groups [9]. When clustering conceptualized bookmarks, the top-down algorithm is more suitable than the bottom-up, because directory path information is already assigned to the bookmarks during conceptualization step.

Instead of ontology, the well-organized web directory services such as Yahoo and Cora can be utilized. In practice, however, these directory services have some drawbacks we have to consider as follows.

- **The multi-attributes of a bookmark.** A bookmark can be involved in more than one concept. As shown in Fig. 2 (1), a bookmark can be included in not only a concept named as `A` but also a concept `B`.
- **The complicated relationships between concepts.** The semantic relationships between concepts can be classified to
  - Redundancy between semantically identical concepts
  - Subordination between semantically dependent concepts.

  In Fig. 2 (1), the concept `A` is a subconcept of the root, but the concept `A` can be redundantly linked as subconcept of the concept `P`. Moreover, the concept `C` can be a subconcept of more than a concept like `P`, as shown in Fig. 2 (2).

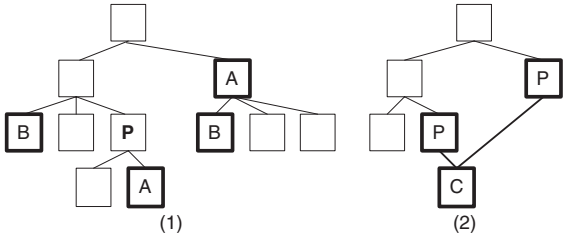

Fig. 2. (1) The multi-attribute of bookmarks; (2) The subordinate relationship between two concepts

When considering influence propagation between concepts, we define notations for semantic analysis dealing with problems caused by web directories. Let the user U_i have the set of bookmarks B_i as follows:

$$B_i = \left\{b_1^i, b_2^i, ..., b_m^i\right\} \tag{1}$$

where m is the total number of bookmarks. To conceptualize B_i, each bookmark in this set is categorized with the corresponding concepts represented as the directory path. Therefore, the set of conceptualized bookmarks C_i is

$$CB_i = \left\{cb_1^i, cb_2^i, ..., cb_n^i\right\}$$
$$CRB_i = \left\{crb_1^i, crb_2^i, ..., crb_a^i\right\}$$
$$C_i = CB_i + CRB_i$$

where n is the total number of concepts including the bookmarks in B_i and a is the number of additional concepts subordinately related with CB_i. Generally, due to the drawbacks of web directories, n becomes larger than m. Here we mention the step for conceptualizing bookmarks by referring to web directories as follows:

```
Function Conceptualization (User)
  var
    counter1, counter2: integer; b: set_bookmark[];
    cb, crb: set_conceptualized_bookmark[];
  begin
    b := Bookmark(User);  counter1 := 1;
    repeat
      cb := cb + Concept(b[counter1]);
      repeat
        counter2 := 1;
        if ((isLinked(Concept(b[counter1]))) = TRUE) then
          crb := crb + Linked(Concept(b[counter1]));
      until counter2 = size(b[counter1])
      counter1 := counter1 + 1;
```

```
    until counter1 = size(b)
    return (cb, crb);
end.
```

The functions `Bookmark` and `Concept` return the set of bookmarks of an input user and the set of concepts matched with an input bookmark by looking up the ontology, respectively. The function `Linked` retrieves the additional concepts related with the input concept. After the function `isLinked` checks if the input parameter is connected from more than one parent concept on the ontology.

3 Extracting User Interests from Conceptualized Bookmarks

In order to extract user interests, the interest map (i-Map) of each user is established and *DOI*s (Degree Of Interest) of the corresponding concepts on the i-Map are measured, according to the following axioms:

Axiom 1. The initial *DOI* of a concept is the number of times that this concept is matched with the set of bookmarks through the function `Conceptualization`. The larger *DOI* of a concept means that the corresponding user is more interested in this concept. This means that this number of times is in linear proportion to user preference for that concept.

$$\text{The Number of Matched Times of Concepts} \propto DOI(C_i)$$

Axiom 2. The *DOI* of a concept is propagated from its subconcepts using this influence propagation:

$$Propagate[DOI(C_i)] = (\log_k(DOI(C_i) + 1))/N \tag{2}$$

where N is the number of total subconcepts of a concept and k is given by

$$k = Variance(DOI(subc(C_i))) + bias = \sigma^2 + bias \tag{3}$$

where $subc(C_i)$ is the set of subconcepts of C_i.

- The dispersion of *DOI*. As the number of subconcepts of a parent is increased, each of them has less influence on its parent concepts.
- The distance between concepts. The closer concepts are more tightly related with each other. In other words, the influence propagation is exponentially increasing, as the distance between concepts becomes closer.

Axiom 3. The *DOI* of a concept is measured from the propagations of all subconcepts and all concepts have influence on the root node.

$$DOI(C_i) = \sum_j [Propagate(DOI(subc(C_i)_j)) \times DOI(subc(C_i)_j)] \tag{4}$$

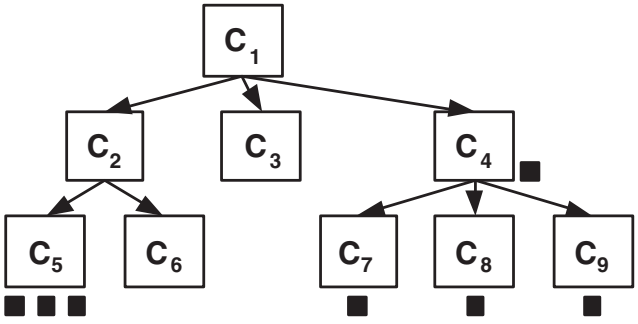

Fig. 3. An example of the conceptualized bookmarks of a user

Axiom 4. Concepts whose *DOI*s are over the predefined threshold value after normalization finally represent user interests.

In Fig. 3, as an example, the black squares indicate the bookmarks of a user, U_i, and assign the initial states, as shown in the following equations:

$$DOI(c_4) = 1, DOI(c_5) = 3, DOI(c_6) = 0,$$
$$DOI(c_7) = 1, DOI(c_8) = 1, DOI(c_9) = 1$$

According to the influence propagation equations, all *DOI*s of other concepts can be computed. The *DOI*s of c_2 and c_4 are as follows:

$$DOI(c_2) = \sum_{k-1}^{2} propagate[DOI(c_k)] \times DOI(c_k) = 1.11$$
$$DOI(c_4) = 1 + (\log_2 2/3 \times 1) \times 3 = 2.0$$

The mean of all *DOI*s is 1.44 and the *DOI* of every concept is assigned after normalization. If the threshold value is 0.2, only c_4 and c_5 are extracted as the most interested concepts for the user. In Fig. 4, the tree represents the user's *i*-Map. Each user is given an *i*-Map, and every time he inserts a bookmark, this *i*-Map is updated.

4 Collaborative Web Browsing with Recommendation

Generally, in computer supported cooperative work (CSCW), a common distinction is made between the temporal and spatial nature of activities. Activities are either co-located or remote and either synchronous or asynchronous [11]. The collaborative web browsing system proposed in this paper is remote and asynchronous because this system is based on a web environment and information about what a participant is interested in extracted from his set of bookmarks and ontology. While browsing to search information, users can be recommended from the facilitator in the following two ways:

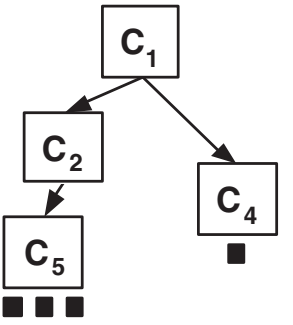

Fig. 4. An *i*-Map representing the high ranked concepts

- By *querying* specific information for the facilitator. After the information about a particular concept is requested, the facilitator can determine who has the maximum *DOI* for that concept by scanning his yellow pages.
- By *broadcasting* new bookmarks of like-minded users from the facilitator. Every time a user inserts a new bookmark, this fact, after conceptualization, is sent to the facilitator. Thereby, users can obtain information related to the common concepts in their own *i*-Map from neighbors.

As shown in Fig. 5, the whole system consists of a facilitator located between the users and the client-side web browser which communicates with the facilitator.

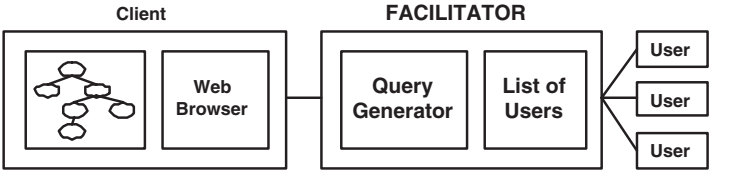

Fig. 5. The whole system architecture

Most importantly, the facilitator must create yellow pages where all users can register themselves. Then, every bookmarking activity can be automatically transmitted to the facilitator.

5 Experimentation

We made up a hierarchical tree structure as a test bed for "Home > Science > Computer Science >" from Yahoo. This tree consisted of about 1300 categories and the maximum depth was eight. For gathering bookmarks, 30 users explored

Yahoo directory pages during for 28 days. Every time users visit a web site related with their own interests, they stored URL information in their bookmark repositories. Finally 2718 bookmarks were collected. In order to evaluate this collaborative web browsing based on extracting user interests, we adopted the measurements *recall* and *precision*. After all of the bookmark sets of the users were reset, these users began to gather bookmarks again after getting the system's recommendations according to their own preferences. During this time, users were being recommended information retrieved from Yahoo based on their interests as extracted up to that moment. As a result, with information recommendations, 80% of the total bookmarks were collected in only 3.8 days, thereby 53.1% of the total time spent previously was saved.

The *precision* was measured by the rate of the inserted bookmarks among the recommended information set. In other words, this was the measurement for the accuracy of predictability. As time passed, the user preferences were changed according to the inserted bookmarks. At the beginning, the precision was especially low because the user preferences were not yet set up. While user interests were being extracted during the first 6 days, the *precision* of recommended information quickly tracked to that of the testing data.

For the rest of the experiment time, the *precision*maintained the same level as that of the testing data because the user interests had already been extracted.

6 Conclusion and Future Work

This paper proposes that bookmarks are the most important evidence to support the extraction of user interests. In order to make up for the structural information, simple URL-based bookmarks were conceptualized by ontology. Then, by establishing an *i*-Map of each user and *DOI* of the concepts on that map, we made it much easier to generate queries for relevant information and to share bookmarks among users. We have implemented a collaborative web browsing system sharing conceptualized bookmarks. Based on the information recommendation of this system, we saved about 53% of the searching time as compared with single web browsing. Moreover, a beginner in a certain field can be helped by finding out valuable hidden information from experts.

As future work, we are considering the privacy problem associated with sharing personal information such as user interests. The visualizing of an *i*-Map is also the next target of this research in order to increase users' intuition recognizing their own preferences quantitatively regarding locations.

Acknowledgement. This work was supported by the Korea Science and Engineering Foundation(KOSEF) through the Northeast Asia e-Logistics Research Center at University of Incheon.

References

1. Maes, P.: Agents that Reduce Work and Information Overload. Comm. of ACM **37**(7) (1994) 31–40
2. Lieberman, H.: Letizia: An Agent That Assists Web Browsing. Proc. of the 4th Int. J. Conf. on Artificial Intelligence (1995) 924–929
3. Armstrong, R., Freitag, T., Mitchell, T.: WebWatcher: A Learning Apprentice for the World Wide Web. AAAI Spring Sym. on Information Gathering from Heterogeneous, Distributed Environments (1997) 6–12
4. Jung, J.J., Yoon, J.-S., Jo, G.-S.: BISAgent: Collaborative Web Browsing through Sharing of Bookmark Information. Proc. of IIP 2000, 16th IFIP World Computer Congress (2000)
5. Lieberman, H., van Dyke, N., Vivacqua, A.: *Let's Browse*: A Collaborative Web Browsing Agent. Proc. of Int. Conf. on Intelligent User Interfaces (1999) 65–68
6. Twidale, M., Nichols, D.: Collaborative Browsing and Visualization of the Search Process. Electronic library and visual information research (1996) 51–60
7. Gruber, T.R.: A translation approach to portable ontologies. Knowledge Acquisition **5**(2)(1993) 199–220
8. Kaufman, L., Rousseeuw, P.: Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley (1990)
9. Maedche, A.: Ontology Learning for the Semantic Web. Kluwer Academic Publishers (2002)
10. Jung, J.J., Jo, G.-S.: Extracting User Interests from Bookmarks on the Web. Proc. of the 7th Pacific-Asia Conf. on Knowledge Discovery and Data Mining (2003) 203–208
11. Rodden, T.: A survey of CSCW systems. Interacting with Computers, **3**(3) (1991) 319–354