

# Deriving Context Specific Information on the Web

Christo Dichev

Department of Computer Science, Winston-Salem State University  
Winston-Salem, N.C. 27110, USA  
dichevc@wssu.edu

Darina Dicheva

Department of Computer Science, Winston-Salem State University  
Winston-Salem, N.C. 27110, USA  
dichevad@wssu.edu

**Abstract:** The Web is huge, unstructured and diverse in quality, which makes searching for information difficult. In practice, few of the documents returned by a search engine are valuable to a user. Which documents are valuable depends on the *context* of the query. Some adequate context information provided in addition to keywords can improve significantly search precision. In this paper we propose a framework for dynamic conceptual clustering of web documents based on clusters of users that share common interests. The basic assumption is that the search results would be more relevant to a user when provided within the context of semantically related documents marked as 'interesting' by a sufficiently large group of users with similar interests. This framework can support personalization of a search based on a search engine that 'knows' the context of the user information needs and uses it to tailor the search results.

## 1 Introduction

The Web is huge and ubiquitous, unstructured, diverse in quality, dynamic and distributed, which makes searching for information principally difficult. General-purpose search engines that use keyword matching are notorious for returning too many matches of little relevance or quality in response to user queries. For example, if you submit the keyword "centroid" to Google almost 60,000 documents will be found. Which documents will be valuable to the user depends on the *context* of the query. The context depends on a number of factors, such as information related to the current request, user's interests, background, education, present professional activities, hobbies, travel and entertainment habits, etc. Search engines, however, treat each request independently from previous requests of the same user and of other web users making similar requests. Therefore the ranked list of documents received in response to the same queries is typically the same and depends neither on the user nor on the context in which the query is made. Some adequate context information provided in addition to keywords can significantly improve the search results. The question is *what type of context information is practical, how to infer that context information and how to use it* for improving search results?

Web users typically search for diverse information. Some searches are sporadic and irregular while others might be related to their interests and have more or less regular nature. An important question is then how to filter out these sporadic, irregular searches and how to combine regular searches into groups identifying topics of interest by observing the user behavior on the web. The fact that a user makes an isolated search for the size of Mars when solving a puzzle does not apparently indicate for any pattern of behavior while regular searches for papers on "Contextual reasoning" are more stable because they identify the user's current interests. Since the causal relations between the user's interests and actions for resource discovery are more stable, the latter are more predictive of user's future behavior. If we are able to identify *topics of interest* for a given user we can infer relevant contextual information associated with that user. Such contextual information when available to search engines could support personalized searches. Our approach to topic identification on the web is based on observations of the searching behavior of large groups of users. The intuition is that a topic of interest can be determined by identifying a collection of web documents that is of common interest to a sufficiently large group of web users.

In the present paper we present a resource discovery framework based on a contextual topology. A problematic point in the original web architecture is that there is no explicit conceptual partitioning of the web

information space. Present web communication assumes mainly *passive information* vs. *active users*. The framework we propose suggests dynamic conceptual partitioning of a portion of web information space into meaningful and more straightforwardly manageable units of information. Such partitioning of the information space would reflect the presence of groups of users that share common interests and possibly some common patterns of behavior. So, it assumes a parallel partitioning of users into groups corresponding to the overlapping objects of interest for each user group. Thus a partitioning of the *user space* into groups of users generates corresponding partitioning of the *web information space* into matching groups of documents. From these user groups we could derive the individual interests of their members to model a context of user's information needs. The partitioning of both spaces is viewed as dynamic and automatic clustering of web documents and users. Since the clustering of documents is based on overlapping interests shared by a sufficiently large group of users, we can assume that such a partitioning reflects the opinion conveyed to the information space by a typical member of each group.

## 2 Our Approach

Keyword queries cannot naturally locate resources relevant to a specific topic. Keywords perform poorly especially in situations in which the search index covers multiple subject areas, as is the case with "Internet" where Google returns 75,000,000 web pages. A promising technique is to guess the context of the user queries. Situations where a search is limited within a group of web documents (a topic) collectively selected by a user and his peers as 'appropriate' illustrate a *context* that is relevant to the current user's information needs. This type of context is also retrospective, because it reflects a portion of the history of the user requests. The fact that all users in a given group like a certain collection of articles is more stable and therefore more predictive than the fact that a particular user likes a given article. We can predict which articles would be of interest to a user based on the articles interesting to the other members of his group. One of the benefits of an explicit representation of a context is that it would enable us to localize the search within a relevant domain and thus to improve the search quality.

Typical search engines can be viewed as 'one size fits all' - all users receive the same responses for the same queries. This model is not always efficient. For example, when searching the web it is not always clear from a request such as "George Harrison" whether the user is looking for the famous "Beatle" or for the owner of the "Menswear of Quality" company. The ambiguity is due to the fact that the context of the search is not generally derivable from the request especially when each request is taken independently from other requests made by the same user. The major questions are: what type of context information is valuable and practical at the same time, how to infer the context information, and how to use this information for improving the search results? The framework for modeling context of user information needs suggested in this paper is based on the observations of documents being viewed (indicated as interesting) by web users. For a given group of users we refer to the collection of documents that are of common interest to all users of that group as a *matching group of documents*. Assume now that on the web there are user groups, from one side, and matching groups of web documents, from another. Then each user searching for information can perceive the web as a personalized list of search results ranked according to the current users' interests. Such a view is inherently *dynamic* as new documents arrive continually and the user groups are dynamic themselves. By synthesizing a conceptual context structure on the web specific to the interests of user groups, this framework provides a ground for a context-based resource discovery. For example, a contextual approach can support personalization of search based on a search engine that knows the contexts of user's information needs and uses that information to tailor the search results. Thus, a request for "George Harrison" may rank links to the owner of "Menswear of Quality" higher than links to the famous pop star for a user interested in stylish dressing.

Our method for topic identification on the web is based on observations of the searching behavior of large groups of web users rather than of a single user. The basic intuition is that a topic of interest can be determined by identifying a collection of web documents (articles, objects) that is of common interest to a sufficiently large group of web users. The assumption is that if a sufficient number of users  $u_1, u_2, \dots, u_m$  driven by their interest are searching independently for a collection of documents  $a_1, a_2, \dots, a_m$ , then this is an evidence that there is a topic of interest shared by all users  $u_1, u_2, \dots, u_m$ . The collection of documents  $a_1, a_2, \dots, a_m$  characterizes the topic of interest associated with that group of users. While the observation on a single user  $u$  who demonstrates interest in objects  $a_1, a_2, \dots, a_m$  is not entirely reliable judgment, the identification of a group of users along with a collection of documents satisfying the relation *interested\_in*( $u_i, a_j$ ) is a more reliable and accurate indicator of an existing topic of interest.

In our approach contexts provide support in two aspects: *personalization* and *community formation*. Personalization refers to both individuals and groups and is based on automatic identification of communities with clustered topical interests. In contrast to directory services such as Yahoo where the web pages are assigned to categories manually, in the suggested framework the notion of context is viewed as a *self-organized* and *dynamic* structure. Self-organization means that the process of identification of existing groups of users and matching groups of documents is based on mining the users web experience for relevant data. Thus contexts are driven by inner dynamics, reflecting the fact that both user groups and matching groups of documents can grow and shrink over the time.

In a practical perspective the proposed approach for identifying a topic of interest is particularly appropriate for *specialized* search engines. First, specialized search engines are focused on finding information within specified fields, for example, Cora (<http://cora.whizbang.com>) is a search engine for computer science research papers. As a result the number of users of specialized search engines is considerably smaller compared to the number of users of general-purpose search engines. Second, specialized search engines use some advanced strategies to retrieve documents. Hence the result list provides typically a good indication of the document content. Therefore, when a user clicks on one of the documents the chances to get relevant information are generally high.

The question is: *how to gather realistic document usability information over some portion of the Web?* One of the most popular ways to get Web usability data is to examine the logs that are saved on servers. A server generates an entry in the log file each time it receives a request from a client. The kinds of data that it logs are: the IP address of the requester; the date and time of the request; the name of the file being requested; and the result of the request. Thus by using log files it is possible to capture rich information on visiting activities, such as who the visitors are and what they are specifically interested in and use it for user-oriented clustering in information retrieval.

The following assumptions provide a ground for the proposed framework. We assume that all users are reliably identifiable across multiple visits to a (search engine) site. We assume further that if a user clicks (saves/selects) a document it is likely that the document is relevant to the query or to the user's current information needs. Another assumption is that all relevant data of user logs are available and that from the large set of user logs we can extract a set of relations of the type: (user\_id, selected\_document). The next step is to derive from the extracted set of relations meaningful collections of documents based on overlapping user interests, that is, to cluster the extracted data set into groups of users with matching groups of documents.

### 3 A Formal Perspective

Given a set of users  $U$ , a set of articles (documents)  $A$  and a binary relation  $uFa$  (user  $u$  is interested in article  $a$ ) determine a pair of subsets  $U_I \in Pow(U)$  and  $A_J \in Pow(A)$  such that

$$U_I = \{u \in U \mid (\forall a \in A_J) uFa\}, \quad A_J = \{a \in A \mid (\forall u \in U_I) uFa\}$$

That is, a topic of interest  $(U_I, A_J)$  is defined by a binary relation  $uFa$  (*interested\_in*( $u, a$ )) and is characterized by the set of all articles  $A_J$  that are common objects of interest to all users in  $U_I$ .

From a formal point of view, the suggested contextual structure on the web can be interpreted as a binary relation between a set of users ( $U$ ) and a set of articles ( $A$ ), called *context*. Thus a context is a triple  $(U, A, F)$ , where  $F \subseteq U \times A$ . Based on an analogy with formal concept analysis (FCA) (Carpineto & Romano 96; Wille 82), a topic of the context  $(U, A, F)$  can be defined to be a pair  $(U_I, A_J)$  where  $U_I \subseteq U$ ,  $A_J \subseteq A$  and,

$$U_I = \{u \in U \mid (\forall a \in A_J) uFa\}, \quad A_J = \{a \in A \mid (\forall u \in U_I) uFa\}$$

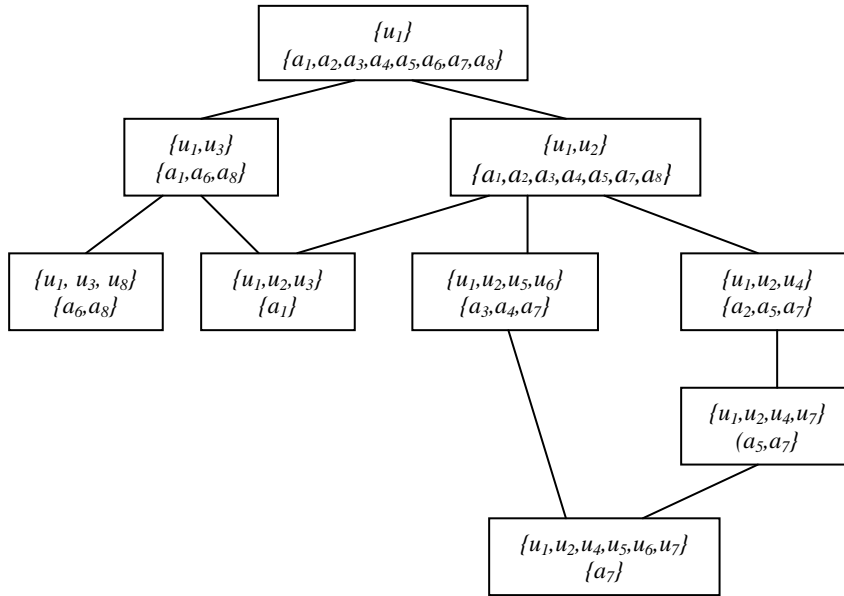
that is  $U_I$  is the set of all users interesting in all articles in  $A_J$  and  $A_J$  is the set of all articles that are common objects of interest to users in  $U_I$ . Exploiting further the analogy with FCA,  $A$  and  $U$  can be interpreted also as a set of *objects* and a set of *descriptors* correspondingly. Then  $A_J$  is the set of all objects possessing all the descriptors in  $U_I$  and conversely  $U_I$  is the set of descriptors held by all objects in  $A_J$ . We may think of the set of articles  $A_u$  associated with a given user  $u \in U$  as represented by a bit vector. Each bit  $i$  corresponds to a possible article  $a_i \in A$  and is on or off depending on whether the user  $u$  is interested in article  $a_i$ . The advantage of this interpretation is that now we can characterize the binary relation between the set of users and the set of articles in terms of *topic lattice*. Let us denote the set of all topics of the context  $(U, A, F)$  by  $T(U, A, F)$ . An ordering relation is easily defined on this set of topics by

$$(U_1, A_1) \leq (U_2, A_2) \leftrightarrow U_1 \subseteq U_2 \text{ or } (U_1, A_1) \leq (U_2, A_2) \leftrightarrow A_1 \supseteq A_2.$$

The set  $T(U,A,F)$  along with the “ $\leq$ ” relation form a partially ordered set that can be characterized by a *concept lattice* (referred here as *topic lattice*). Each node of the topic lattice is a pair composed of a subset of articles and a subset of corresponding users. In each pair the subset of users contains just the users sharing common interest to the subset of articles and similarly the subset of articles contains just the articles sharing overlapping interest from the matching subset of users. The set of pairs is ordered by the standard “set inclusion” relation applied to the set of articles and to the set of users that describe each pair. The partially ordered set can be represented by a Hasse diagram, in which an edge connects two nodes if and only if they are comparable and there is no other node - intermediate topic in the lattice, i.e. each topic is linked to its maximally specific more general topics and to its maximally general more specific topics. The ascending paths represent the subclass/superclass relation. The bottom topic is defined by the set of all users; the top topic is defined by all articles and the users (possibly none) sharing common interest in them. A simple example of user and information spaces is presented in (Tab. 1). The corresponding lattice is presented in (Fig. 1).

	$a_1$	$a_2$	$a_3$	$a_4$	$a_5$	$a_6$	$a_7$	$a_8$	.
$u_1$	1	1	1	1	1	1	1	1	.
$u_2$	1	1	1	1	1	0	1	0	.
$u_3$	1	0	0	0	0	1	0	1	.
$u_4$	0	1	0	0	1	0	1	0	.
$u_5$	0	0	1	1	0	0	1	0	.
$u_6$	1	0	1	1	0	0	1	0	.
$u_7$	0	0	0	0	1	0	1	0	.
$u_8$	0	0	0	0	0	1	0	1	.
.	.	.	.	.	.	.	.	.	.

**Table 1:** An example of user and information spaces.



**Figure 1:** A lattice corresponding to the user and information spaces presented in Table 1.

In contrast to conceptual clustering (see Balabanovich & Shoham 97) where the descriptors are static, in the suggested approach the users who play a role of descriptors are *dynamic*: in general, a user’s interest can not be specified completely and his topical interests change over time. Hence, the lattice describing the topical structure is dynamic too. This induces some results based on the following assumptions. A collection of articles

$A_j$  from an existing topic  $(U_i, A_j)$  can only be expanded. This is implied by the conjecture that documents, which are interesting for a user  $u$  remain interesting to him. Therefore, an expansion of the collection of articles with respect to a topic  $(U_i, A_j)$  will not impose any change of existing links. Indeed, an expansion of  $A_j$  to  $A'_j$  results in an expansion of all parent (descendent) collections  $A_m, A_n$ , such that  $A_j \subseteq A_m \subseteq A_n$ , i.e. from  $A_j \subseteq A'_j \rightarrow A_m \subseteq A'_m$  and therefore  $(U_n, A_n) \leq (U_m, A_m) \rightarrow (U_n, A'_n) \leq (U_m, A'_m)$ . Analogous relations hold with ancestor nodes. That is, an expansion of an existing collection of articles preserves the structure of the lattice. The next assumption is a formal support of our intuition that the search domain relevant to the user  $u \in U_i$  includes a subset of articles to which other members of the group  $U_i$  have demonstrated interest. These are collections of articles  $A_k$  of the topics  $(U_m, A_k)$ , such that  $u \notin U_m \cap U_i \neq \emptyset$ .

One of the main factors in a page ranking strategy involves the location and frequency of keywords in a web page. Another factor is link popularity - the total number of Web sites that link to a given page. However, present page rank algorithms typically do not take into account the current user and specifically user's interests. Assume that we have partitioned users into groups associated with their topics of interest (as collections of documents). A modified ranking algorithm can be obtained by extending the present strategy with an additional factor involving the number of links to and from a topic associated with a given user. In this case the page ranking strategy takes into consideration user's interest encoded in the number and the levels of links to a topic associated with a given user. So, for a user  $u \in U_i$ , where  $(U_i, A_j)$  is a topic, the page rank of an article  $a$  should depend on the linkage structure to the articles  $a_i \in A_j$  representing the topic of interest of user  $u$ . We can interpret a link from page A to page B as a vote of page A for page B. Thus votes cast by pages that are from the users topic weigh more heavily and help to make other pages 'more-important'. This strategy makes page-ranking *user oriented*. Such a strategy promotes pages related to users' topics of interests. From an "active users" perspective this approach enables us to recognize a community of users for which a given article is most likely to be interesting.

This approach suggests also an ordering relation ( $\prec$ ) for ranking articles returned in response to a request from a user  $u$ , assuming that  $U_i$  is the maximally general group such that  $u \in U_i$ . Thus  $a_1 \prec a_2$  if there exist collections  $A_1$  and  $A_2$ ,  $a_1 \in A_1$ ,  $a_2 \in A_2$ , where  $A_1$  and  $A_2$  are components of existing nodes  $(U_i, A_1)$  and  $(U_2, A_2)$ , such that  $U_1 \cap U_i \neq \emptyset$ ,  $U_2 \cap U_i \neq \emptyset$  and  $|U_1| < |U_2|$ , i.e. the more members of the group  $U_i$  have expressed an interest in a given article the better. The later ordering relation can be incorporated into the suggested ranking method.

An important characteristic of the method is that it does not require explicit representation of the Web objects, due to the fact that it exploits "membership" relations. The set of objects  $A_u$  are identified based on their relevance to the context "interesting to the user  $u$ ", rather than on specific syntactic properties of their representations. Therefore it can cover objects behind the conventional search forms, such as pdf files, images, music files, and compressed archives. In many cases the user is not certain of what information exactly to look for and needs to learn more about the content of the information space. In such cases browsing a collection of documents generated on the base of the current context can be a good navigational strategy. Thus this framework supports information retrieval based on contextual browsing, where user  $u \in U_i$  can navigate through the matching collection of articles  $A_j$ .

## 4 Related Works and Conclusion

The Web is probably the richest information repository in human history, but it is usually hard and time consuming to find desired information there. The low precision of the web search engines due to the lack of contextual knowledge makes it difficult to find relevant information. The focus of the current efforts of the Web research community is mainly on optimizing the search, assuming active users vs. passive information.

Recently there has been much interest in supporting web users through collecting web pages related to a particular topic (Brin & Page 98), (Chakrabarti et al. 99), (Mukherjea 00). These approaches aimed at topic specific resource discovery typically exploit connectivity for topic identification but not community identification. Community identification does not play any significant roles in these methods and therefore user search experience within a community is ignored. Some systems such as (Kumar et al. 99) do exploit the experience of other web surfers to derive clustered topical interests but the focus is on organizing surfing history in coherent topics for later use. The problem of identifying community structure on the Web was addressed in (Kumar et al. 99). However, the approach employed for community identification is based on analysis of the Web graph structure and is not explicitly related to resource discovery. In collaborative filtering systems (see Balabanovich & Shoham 97) items are recommended on the basis of user similarity rather than

object similarity. Each target user is associated with a set of nearest neighbor users (by comparing their profiles) who act as ‘recommendation partners’. These systems are aimed at recommending items from a fixed topic/database. In contrast, our approach is aimed at large scale resource discovery based on derived topics reflecting similarity of interests among users. A derived benefit of such an approach is localizing the search within an individual topic of interest. Categories and formal concept analysis (Carpineto & Romano 96), (Wille 82), modeling and using contexts (Dichev et al. 01), (Lawrence 00), (Glover et al. 99) have been studied for a long time, motivated by the need for a formalization of the notions of concept and context. A major practical issue is the level at which contexts are defined and analyzed. The context information used in Inquirus 2 metasearch engine (see Glover et al. 99) is in the form of category of the desired document. The context is used to select the search engines, to modify queries and to select the ranking strategy. Our framework is close in spirit to the application of Galois’ concept lattices (Carpineto & Romano 96), but the grouping of web objects into classes is based on dynamic descriptors associated with web users.

In this paper we have presented a novel framework for information retrieval on the web. The basic assumption is that the results of a search would be more relevant to a given user when provided within the context of semantically related objects marked as “interesting” by their peers. In addition, a contextual structure would help users in getting better insight about the scope of their search, localizing the search, reducing the amount of time for deciding on the relevance of the hits of a search, comparing the hits to objects selected by their peers, optimizing the search strategy, etc. An additional advantage of the presented approach is that the topical clustering can cover objects behind the conventional search forms including non-indexed web objects, such as pdf files, images, music files, and compressed archives. By partitioning web users into groups it would be possible to make useful predictions about some other shared features and patterns of users’ behavior and derive some correlating properties regarding the web information space. Moreover, the identification of significant groups of users with similar interests can help producers of information to reach interested consumers in a timely manner.

## References

- Balabanovich, M., & Shoham Y. (1997). Content-based Collaborative Recommendation. *Communications of the ACM* 40(3), 66-72.
- Brin, S., & Page, L. (1998). The Anatomy of a Large-scale Hypertextual Web Search Engine. In *7<sup>th</sup> Int. WWW Conf.*, Vol. 7. Online at <http://google.stanford.edu/backrub/google.html>.
- Carpineto, C., & Romano, G. (1996). A Lattice Conceptual Clustering System and Its Application to Browsing Retrieval. *Machine Learning* 24, 95-122.
- Chakrabarti, S., van den Berg, M., & Dom, B. (1999). Focussed Crawling: a New Approach to Topic-specific Web Resource Discovery. In *Proceedings of the Eight International World Wide Web Conference*, Toronto, Canada, 545-562.
- Dichev, C., Dicheva, D., & Radenski, A. (2001). A Framework for Dynamic Topic Clustering on the Web. In: Proc. of *The International Conference on Internet Computing (IC'2001)*, Las Vegas, Nevada, 885- 993.
- Glover, E., Lawrence, S., Birmingham, W., & Gilles, L. (1999). Architecture of a Meta-search Engine that Supports User Information Needs. In Proc. of *CIKM 99*, Kansas City, 210-216.
- Kumar, R., Raghavan, P., Rajagopalan, S., & Tomkins, A. (1999). Trawling the Web for Emerging Cyber-communities. In *Proc. of the Eight Int. World Wide Web Conference*, Toronto, 403-415.
- Lawrence, S. (2000). Context in Web Search. *IEEE Data Engineering Bulletin*, 23(3), 25-32.
- Mukherjea, S. (2000). WTMS: A System for Collecting and Analyzing Topic Specific Web Information. In *Proceedings of the Ninth International World Wide Web Conference*, Amsterdam.
- Wille, R. (1982). Restructuring Lattice Theory: An Approach Based on Hierarchies of Concepts. In: I. Rival (ed.): *Ordered Sets*. Reidel, Dordrecht-Boston, 445-470.