Query Driven Conceptual Browsing: A Semi-Automated Approach for Building and Exploring Concepts on the Web

Kinshuk Jerath and Balaji Padmanabhan
Operations and Information Management Department
The Wharton School, University of Pennsylvania
{kinshuk, balaji}@wharton.upenn.edu

Abstract

The presence of communities, which are groups of highly cross referenced pages together representing a single concept, is a striking feature of the World Wide Web. Quite often a group of communities, each topically coherent within itself, may be related through a common concept manifested in each of them. Motivated by this observation, we present a method for *query-driven conceptual browsing* for exploring concepts on the Web starting from a user-specified query. We show how this idea is related to prior work on learning concept maps and on Web Mining, and discuss the application of conceptual browsing for user-driven exploration and discovery of new concepts on the Web.

1. Introduction

Despite its decentralized growth, the Web has been found to have neat structure. Web pages are often found to be organized into communities which are groups of highly cross referenced pages about specific concepts. Our work hinges on the observation that several communities, each about a specific concept, often have another common concept connecting them. For instance, communities on the three different concepts of *cartography*, *volcanology* and *oceanography* all have *earth science* as a concept that connects them, since each of these is an *earth science*. Motivated by this observation, we present a method for *query-driven conceptual browsing* (QDCB) for exploring concepts on the Web starting from a user-specified query.

QDCB is an iterative process with three parts[1]. First, starting from a user-specified query, treated as an initial concept, we learn different communities by clustering the pages returned for a query using a graph clustering technique [3]. This step directly uses prior work on learning communities on the Web [7, 8]. Second, using a semi-automated approach we determine the concepts corresponding to the different communities and these concepts are viewed as the results of the Web search. Third, we let the user pick a specific returned concept, formulate a query based on the chosen concept and repeat the process with this new query, thereby learning new related concepts, potentially very different from those returned for the original query. Building on prior work on concept spaces in IR and on work on exploiting the link structure of the Web and learning communities, in this paper we discuss the application of conceptual browsing for user-driven exploration and discovery of new concepts on the Web through this chain-like learning. We propose an algorithm for the same and implement it on search results derived from using Google's Web Services-based API and report results.

2. Overview of the Approach and Related Work

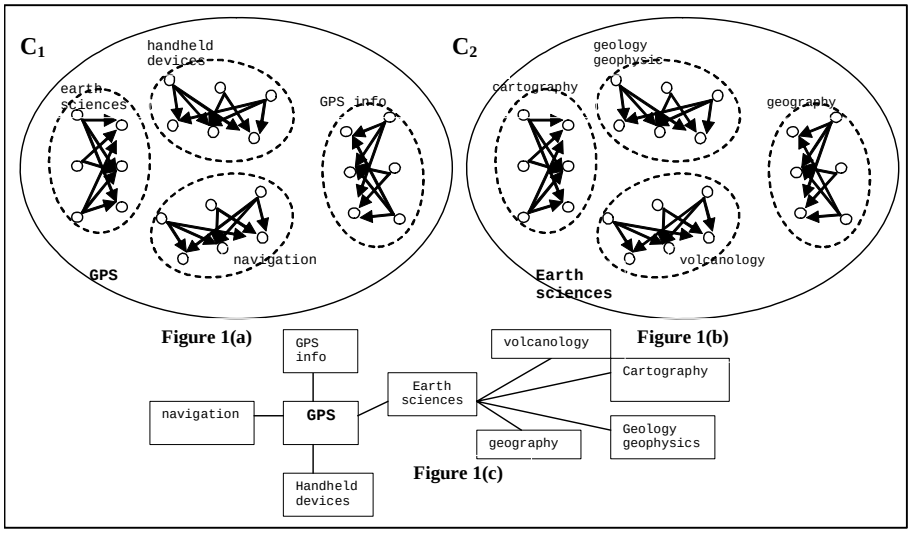
The idea of conceptual browsing is that by presenting the user with a set of concepts related to a query, the Web may be browsed one concept at a time rather than one page at a time. Further this can be used to

[1] The reader should keep in mind that by definition, a *community* is topically coherent and therefore about one *concept*. We assume that a concept can be represented using a few *keywords*. Thus a community is about one concept and can be represented by a few keywords. A *query* is a set of words which when fed into a search engine, returns Web pages from the community about a certain concept on the Web. The notion of a query for a concept is thus akin to that of keywords for the concept, except that it may be fine tuned as per the requirements and syntax of a specific search engine.

learn new related concepts in a semi-automated manner. Below we describe this based on an example and then discuss how this relates to other work.

We implemented the QDCB approach for the query "GPS" (using Google's Web Services-based API to generate the pages that we processed) and our results are in Figure 1. As shown in the graph in Figure 1(a) the set of returned results C_1 on "GPS" contains sub-clusters C_{11}, C_{12}, C_{13}, C_{14} (obtained using a graph clustering technique [3]). Based on a semi-automated approach (described in the next section) we characterize the concepts represented by these clusters as *earth sciences*, *handheld devices*, *general information on GPS* and *navigation*. These concepts represent independent, yet related clusters: GPS devices are indispensable for field research in the earth sciences, pages about handheld devices are relevant since many of these devices contain GPS chipsets, there is a community of pages with general information on GPS technology and GPS is used for navigation. Note that these four groups are about four different concepts, yet have a common concept (GPS) that connects them.

Above we illustrated the first two parts of our approach – clustering and characterizing the concept of each cluster. Next we illustrate concept expansion, where we learn new concepts that are linked to the previously discovered ones. Assume the user picks *earth sciences* as the concept to expand further. The above procedure is carried out again with *earth sciences* as the new query. Note that how exactly a new query is formed based on the chosen concept is an important question in itself, and we address this in the next section. Based on the new query we obtain the set C_2 which is again partitioned into sub-clusters C_{21}, C_{22}, C_{23}, C_{24} as shown in Figure 1(b). As the results show, the new clusters learned are about various concepts in earth sciences, such as *cartography*, *geology*, *volcanology* and *geography*. In this example it is also useful to note that: (a) the original earth sciences cluster of pages from C_1 is now a subset of C_2, which also has many more pages that were not part of C_1 (since the expansion query "earth sciences" is different from the original query "GPS") and (b) the pages in the original earth sciences cluster are distributed across all the four clusters in C_2. It may be possible to constrain the expansion such that the earth sciences cluster is entirely part of one specific sub-cluster in C_2. In this paper we do not address how to do this, although in many other concepts that were expanded we observed that this naturally occurs.



Finally note that these new clusters are related to the clusters under GPS in a chain-like form with *earth sciences* as the bridging link. The user is, in effect, conceptually browsing a chain of concepts shown in Figure 1(c). Alternatively, the user can choose to advance in a different direction by extending on the concepts *navigation* or *handheld devices* each of which would further lead to its own set of sub-clusters and thus help the user learn new conceptual chains.

Browsing concepts as described above is closely related to prior work on concept spaces in IR [2] and the more recent use of topic maps [6] that is relevant in the context of the semantic web [1]. Concept spaces in IR referred to a thesaurus-like description of concepts and their relationships. More recently

topic maps [6] refer to a standard that can be used to define topics and relationships between topics. In both cases, the idea is that these maps can aid in the information retrieval process. Our approach can be viewed as one specific method for learning parts of concept spaces for documents on the Web. When viewed in this context there are three small, but important differences in our QDCB approach: (i) we present a method for learning concepts based on using the hubs and authorities framework in [7] for the Web (ii) we do not learn complete ontologies, concept spaces or topic maps in a given domain, but just learn fragments of these that are related to a user query – a less ambitious approach, but one that is less complex as well, and (iii) we describe a process where the concept space idea is used in conjunction with carefully constructed user-query expansion in order to help users browse a chain of concepts on the Web.

In addition to the above work, there is also work on visualizing the results of Web searches graphically in order to help users make sense of the retrieved information. For example, Kartoo (www.kartoo.com) and Grokker (www.grokker.com) are two approaches that display Web search results as a graph. There has also been research on aggregating and visualizing search results [4, 9], i.e. presenting the search results from a web search engine as a set of clusters, each on a specific topic. Also recently [10] present a method for visualizing Web search graphically as concepts, but their approach relies on having an ontology and metadata in documents. While the commercial tools do not have published algorithms that describe the procedure used, our approach is different in that we do not use any prior background knowledge in this process, present a specific way to build these concepts using hubs and authorities, and present a method that enables concept expansion - which is not handled in these current systems yet. However these efforts underscore the observation that aggregation of search results and graphical browsing such as described here can be useful and that methods for systematically doing so are needed. In the next section we formalize the QDCB approach outlined here and describe an algorithm for query-driven conceptual browsing.

3. Algorithm QDCB

In this section we present an algorithm for query-driven conceptual browsing. The input to the algorithm is a query *q* and the maximum recursion depth MAX_DEPTH. We start by searching the Web using the query *q*. Next we construct the graph representation of the result set and partition this graph into sub-clusters, each representing a concept. These concepts are then characterized by a set of descriptions and keywords. For each concept an expansion query is built using the keywords, and the expansion step is carried out by calling the procedure recursively on this query. The chains between concepts are established when the recursion stops. The algorithm is outlined in Figure 2 and is explained in detail below.

```
ConceptualBrowsing(q, MAX_DEPTH)
       FindConcepts(q, 0, MAX_DEPTH)

FindConcepts(q, d, MAX_DEPTH)
     1.  Conduct a Web search using the query q and construct the rootset
     2.  Crawl all outlinks and inlinks to the rootset and construct the baseset
     3.  Represent the baseset as a graph G and using the spectral graph partitioning
         algorithm make clusters C1, C2, …Ck
     4.  For each cluster Ci
           I.  Find the top m authorities and hubs
          II.  Using the authorities and hubs characterize Ci by the description di and a
               keyword set Ki
         III.  Construct the query qi for expansion
          IV.  If d < MAX_DEPTH execute FindConcepts(qi, d+1, MAX_DEPTH)
```

Figure 2: Algorithm QDCB

In order to avoid having graphs with a very large number of nodes to cluster we adopt the approach used previously in [7] to build communities. First, using the query *q* we conduct a Web search and crawl the top *n* pages returned, also called the *rootset R*. As argued in [7], the main pages within a community are the *hubs* ("directory" pages which are collections of important pages) and *authorities* (the "authoritative" pages for any community). These pages, while important for characterizing communities, may not be part of *R*. However, these are expected to be linked to or linked from at least a few pages in *R*.

The idea [7] therefore is to expand the rootset to include all outgoing links and incoming links for all pages in it. This expanded set is termed the *baseset B.*

We next construct the graph representation *G* of the set *B*, treating each page as a node and each hyperlink as an edge in the graph. Using the spectral graph partitioning algorithm in [3] we learn sub-clusters $C_1, C_2 \ldots C_k$. We now need a description and a set of keywords for each of the clusters obtained. Constructing these is a subjective issue and we therefore do this in a semi-automated manner. First, we find the top *m* hubs and authorities for each cluster and use them as representative pages for the clusters. From these we manually generate a description and keywords for the clusters. We next use these keywords to construct an expansion query for each cluster. A key issue here is the choice of this query (call it *seed query*) that is used to expand a particular sub-concept (call it *seed concept*). The query should be i) specific enough to include a large fraction of the seed set, so that the fresh crawl is related to the seed concept ii) general enough to help discover new concepts iii) such that the expansion from the seed set stays clear of the seed sets for other sub-concepts, to ensure that this learns concepts different from other sub-concepts

Based on our experience with several runs of QDCB, from these representative pages and keywords of the pages an expansion query for each cluster is easy to determine. In general QDCB provides a framework within which a user will play an active role in concept expansion and discovery. Formulating expansion queries is a key part of this, although in future we will study methods to provide users automatically with a set of expansion queries to consider. See table 1 for example expansion queries. We thus have descriptions $d_1, d_2, \ldots, d_k$, sets of keywords $K_1, K_2, \ldots, K_k$ and expansion queries $q_1, q_2, \ldots, q_k$. Using the expansion query for each concept we repeat the process to learn new concepts.

4. Implementation and Results

We implemented the algorithm for various queries and in this section we describe additional results for the query "GPS". To obtain the pages to build the graphs we use Google's Web Services-based API which gives query access to Google's Web search facility and its cache of over 4 billion Web pages through Java programs. Google Web API support the same search syntax as the www.google.com site. For any query we crawl the first 150 results returned by Google into the rootset. For each page in the rootset we then crawl, again using the Google API, all the pages which link to it. Next we parse the rootset to obtain all outlinks and crawl them. We, however, do not crawl hyperlinks between two pages in the same website, since these are mostly navigational links.

The description, keyword set and expansion query are constructed for each sub-cluster using the top 30 hubs and authorities[2]. We ensure minimal overlap with other seed concepts by imposing that the set of pages returned by the seed query does not have the keywords for the other sub-concepts by constructing the appropriate Google query according to the syntax specified in the Google Web API documentation.

For the query "GPS" some of the clusters learned are small, with very few pages in them and these are ignored. As mentioned in Section 2, the four main clusters learned are about *GPS information*, *navigation systems, earth sciences* and *handheld devices.* The algorithm is called recursively once to expand and obtain sub-clusters for each of the three topics *navigation systems, earth sciences* and *handheld devices*. The results have been shown in the tables 1 and 2. The chains of concepts learned have been shown in Figure 3. As seen from Table 2 and Figure 3, we learn several new concepts related to GPS and a majority of them appear to make sense[3]. For lack of space we do not discuss the concepts in detail

[2] It is noteworthy here that hub pages contribute to the generality of the concept description while authorities contribute to the specificity. For instance, for the cluster *"handheld devices with GPS functionality"* the hubs link to various pages on handheld devices with and without GPS functionality, while the authorities mostly are about handheld devices with GPS technology embedded in them.

[3] There were occasional exceptions. For instance we found the cluster on literary classics strange. On further investigation we found that this was due to a single page (www.pnavy.com) returned by Google in the rootset for "navigation", which had several links to literary classics.

here. More generally, Figure 3 provides an example of what a user may get from our approach after several query and concept expansions.

Topic	Representative Pages	Description and Keywords	Expansion Query
GPS Info	www.suomensotilas.fi/nettisotilas/Lehti/NS_GPS7.html www.cetusgps.dk/links.html geospatial.osu.edu/resources/handheldgps.html	Information about various aspects of GPS	*Did not expand*
Earth Sciences	dir.i-une.com/Science/Earth_Sciences/Geomatics/ directory.google.dk/Top/Shopping/Publications/Maps/	Geology, Geomatics, Geodesy, Cartography; "earth sciences"	"earth sciences" – "handheld devices" – navigation
Handheld Devices	www.garmin.com/mobile/ www.pocketpcminds.com/	Handheld devices with GPS and other functionality; "handheld devices"	"handheld devices" – "earth sciences" – navigation
GPS Navign.	www.travelbygps.com/ www.navtechgps.com/ www.waypoints.de/	Navigation systems especially using GPS and satellite technology	Navigation – "handheld devices" – "earth sciences"

Table 1. Initial clusters for query "GPS", cluster descriptions and the chosen expansion queries

Sub cluster name	Clusters found after expansion
Earth Sciences	1. Volcanology 2. Research Institutes in Earth Sciences 3. Various earth sciences like geology, geophysics, geochemistry 4. Geography, oceanography 5. Cartography
Handheld Devices	1. Media esp. related to IT sector e.g. news websites 2. Mobile devices in general esp. cellphones 3. Information and review websites for IT products esp. mobile devices
GPS Navigation	1. Aviation 2. Outdoor camping 3. Boating and navigation in water 4. Cluster on literary classics and creative writing

Table 2. New concepts derived from the expansion step

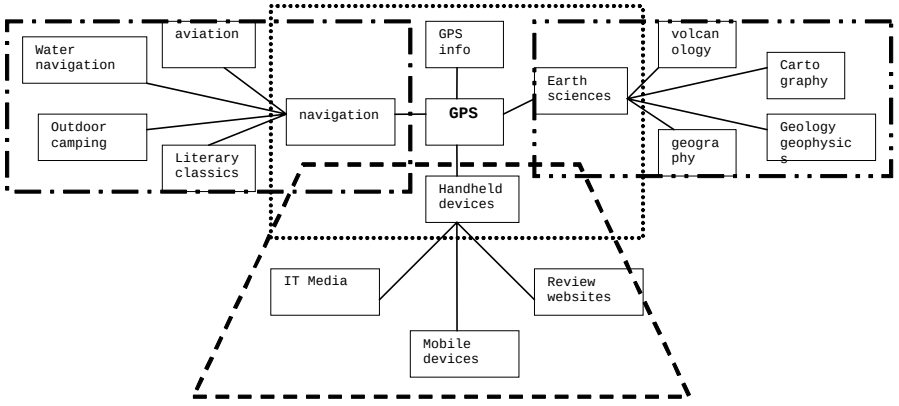


Figure 3: Chains of concepts learned starting from "GPS" after one level of recursion

5. Discussion

In this paper we presented the idea of query-driven conceptual browsing which may be thought of as an exploration tool that helps a user explore a chain of concepts on the Web. We presented a semi-automated method which learns such chains from a given query and presented results from applying this to search results derived from Google's Web Services based API. To make the approach practical, the step involving query construction for new clusters needs to be automated. We can use one of several available approaches for this. [5] presents a hierarchical inferential learning method to produce concise keyword

descriptions for documents. [11] presents a method which takes partially structured source text, extracts information content from it, and presents the most important content in a manner sensitive to the needs of the user and the task. These approaches can be easily extended to work on clusters of documents to produce the descriptions, keywords and queries that we need to automate QDCB.

While the initial results are encouraging, we also need to systematically evaluate the approach. This can be done as follows. For the semi-automated approach user experiments can be done to compare the effectiveness of QDCB. For instance a user can be asked to explore concept chains using QDCB on a set of chosen starting queries and to contrast this experience with topic exploration using a standard search engine. If using a fully automated system, the topic maps obtained from QDCB can be evaluated by comparing them to known topic maps from previous expert knowledge.

There are several potential business and policy applications for the approach presented here. First, the idea of conceptual browsing can be used to develop (or improve) a search engine to enable user-driven concept discovery. The search engine Teoma provides a mechanism to narrow a user's initial search. A natural extension of such an approach is concept expansion and the method described here may be one approach for expansion. Second, the Web is an important tool for business analysts today given the wealth of information on various industries and applications. Our method can be a valuable tool for such analysts to learn a holistic view of industry structure based on relationships on the Web and also to understand new applications or technologies at a higher level of abstraction. For example, we observe that a number of pages on petroleum exploration appear in the sub-cluster on *geology* which is related to *GPS*. This suggests exploring the possibility of using GPS technology in oil exploration, which is in fact an upcoming use of GPS technology these days. Third, policy analysts can learn about issues related to a certain topic like abortion or pollution at a higher level of abstraction based on our approach. This may aid in understanding some of the relevant issues and perhaps provide some input for policy development. Fourth, our approach can be used to learn a chain of concepts related to a specific application or product, and important pages on related concepts may be useful for advertising. For example, based on the results, GPS devices may be advertised in the hubs and authorities for oil exploration.

References

1. Berners-Lee, T., Hendler, J., Lassila, O. *The Semantic Web*. Scientific American 184 (2001) pp.34-43
2. Chen, H., Lynch, K.J. *Automatic Construction of Networks of Concepts Characterizing Document Databases.* IEEE Trans. on Systems, Man, and Cybernetics 22(5):pp. 885-902, 1992.
3. Ding, C. H. Q., He, X., Zha, H. *A spectral method to separate disconnected and nearly-disconnected Web graph components.* KDD 2001: 275-280, 2001.
4. Cutting, D., Karger, D., Pedersen, J., Tukey, J. W. *Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections.* Proc. of the 15th Annual Int. ACM/SIGIR Conf. 1992.
5. Harik, G., Shazeer, N. M. *Method and Apparatus for Probabilistic Hierarchical Inferential Learner.* United States Patent Application No. 20040068697.
6. http://www.topicmaps.org/xtm/1.0/
7. Kleinberg, J. M. *Authoritative Sources in a Hyperlinked Environment.* Journal of the ACM, Vol. 46, No. 5, pp. 604-632, 1999.
8. Kumar, S.R., Raghavan, P., Rajagopalan, S., Tomkins, A. *Trawling the Web for emerging cyber-communities.* In Procs.of WWW8, Toronto, Canada, May 1999.
9. Kummamuru, K., Lotlikar, R., Roy, S., Singal, K., Krishnapuram, R. *A hierarchical monothetic document clustering algorithm for summarization and browsing search results.* In Procs. of WWW 2004: 658-665
10. Kunz, C., Botsch, V. *Visual Representation and Contextualization of Search Results – List and Matrix Browser.* Proc. of Int. Conf. on Dublin Core and Metadata for e-Communities, 2002.
11. Schiffman, B., Mani, I., Concepcion, K.J. *Producing Biographical Summaries: Combining Linguistic Knowledge with Corpus Statistics.* Proc. European Association for Computational Linguistics, 2001.