Exploratory Web Searching with Dynamic Taxonomies, Results Clustering and Visualization

Panagiotis Papadakos

1 Institute of Computer Science, FORTH-ICS, GREECE, 
2 Computer Science Department, University of Crete, GREECE 
papadako@ics.forth.gr

Abstract. The general objective of this dissertation is to elaborate on methods and techniques for exploring and progressively refining large volumes of heterogeneous information, through the provision of appropriate models of interaction and visualization techniques. The goal is to support advanced in accuracy and completeness answers, appropriate for supporting decision making, while the user is not obliged to formulate complex queries or to use specific user profiles. Instead, it will be attempted to provide user with "all" beneficial options to adjust or restrict the information received, in a summarized, concise and intuitive manner. The navigation should provide appropriate data visualization techniques, in order to exploit the ability of the human brain for rapid understanding and perception of visual information, and the ability to discover standards and relationships through it.

1 Introduction

Web Search Engines (WSEs) typically return a ranked list of documents that are relevant to the query submitted by the user. For each document, its title, URL and snippet (fragment of the text that contains keywords of the query) are usually presented. It is observed that most users are impatient and look only at the first results [1]. Consequently, when either the documents with the intended (by the user) meaning of the query words are not in the first pages, or there are a few dotted in various ranks (and probably different result pages), it is difficult for the user to find the information he really wants. The problem becomes harder if the user cannot guess additional words for restricting his query, or the additional words the user chooses are not the right ones for restricting the result set.

One solution to these problems is results clustering [33] which provides a quick overview of the search results. It aims at grouping the results into topics, called clusters, with predictive names (labels), aiding the user to locate quickly documents that otherwise he wouldn’t practically find especially if these documents are low ranked (and thus not in first result pages). Another solution is to exploit the various metadata that are available to WSEs (like domain, dates, language, filetype, etc). Such metadata are usually exploited through the advanced search facilities that some WSEs offer, but users very rarely use these
services. A more flexible and promising approach is to exploit such metadata in the context of the interaction paradigm of faceted and dynamic taxonomies [23, 27, 24], a paradigm that is used more and more nowadays. Its main benefit is that it shows only those terms of the taxonomy that lead to non-empty answer sets, and the user can gradually restrict his focus using several criteria by clicking. In addition this paradigm allows users to switch easily between searching and browsing.

The rest of this paper is organized as follows. Section 2 discusses requirements, related work and background information. Section 3 describes our vision for exploratory search systems. Section 4 describes implementation and reports preliminary experimental results while Section 5 offers an overview of the proposed approach.

2 Requirements & Background

**Results Clustering** Results clustering algorithms should satisfy several requirements. First of all, the generated clusters should be characterized from high intra-cluster similarity. Moreover, results clustering algorithms should be efficient and scalable since clustering is an online task and the size of the retrieved document set can vary. Usually only the top $- C$ documents are clustered in order to increase performance. In addition, the presentation of each cluster should be concise and accurate to allow users to detect what they need quickly.

Cluster labeling is the task of deriving readable and meaningful (single-word or multiple-word) names for clusters, in order to help the user to recognize the clusters/topics he is interested in. Such labels must be predictive, descriptive, concise and syntactically correct. Finally, it should be possible to provide high quality clusters based on small document snippets rather than the whole documents.

In general, clustering can be applied either to the original documents (like in [5, 10]), or to their (query-dependent) snippets (as in [33, 26, 7, 34, 9, 28]). Clustering meta-search engines (e.g. clusty.com) use the results of one or more WSEs, in order to increase coverage/relevance. Therefore, meta-search engines have direct access only to the snippets returned by the queried WSEs. Clustering the snippets rather than the whole documents makes clustering algorithms faster. Some clustering algorithms [7, 6, 31] use internal or external sources of knowledge like Web directories (e.g. DMoz3), Web dictionaries (e.g. WordNet) and thesauri, online encyclopedias and other online knowledge bases. These external sources are exploited to identify significant words/phrases, that represent the contents of the retrieved documents or can be enriched, in order to optimize the clustering and improve the quality of cluster labels.

One very efficient and effective approach is the **Suffix Tree Clustering (STC)** [33] where search results (mainly snippets) can be clustered fast (in linear time), incrementally, and each cluster is labeled with a phrase. Overall and for the problem at hand, we consider important the requirements of relevance, browsable

---

3 www.dmoz.org
summaries, overlap, snippet-tolerance, speed and incrementality as described in [33]. Several variations of STC have emerged recently (e.g. [4, 13, 28]).

**Exploratory Search and Information Thinning** Most WSEs are appropriate for focalized search, i.e. they make the assumption that users can accurately describe their information need using a small sequence of terms. However, as several user studies have shown, this is not the case. A high percentage of search tasks are exploratory [1], the user does not know accurately his information need, the user provides 2-5 words, and focalized search very commonly leads to inadequate interactions and poor results. Unfortunately, available UIs do not aid the user in query formulation, and do not provide any exploration services. The returned answers are simple ranked lists of results, with no organization.

We believe that modern WSEs should guide users in exploring the information space. **Dynamic taxonomies** [23] (faceted or not) is a general knowledge management model based on a multidimensional classification of heterogeneous data objects and is used to explore and browse complex information bases in a guided, yet unconstrained way through a visual interface. Features of faceted metadata search include (a) display of current results in multiple categorization schemes (facets) (e.g. based on metadata terms, such as size or date), (b) display categories leading to non-empty results, and (c) display of the count of the indexed objects of each category (i.e. the number of results the user will get by selecting this category). An example of the idea assuming only one facet, is shown in Figure 1. Figure 1(a) shows a taxonomy and 8 indexed objects (1-8). Figure 1(b) shows the dynamic taxonomy if we restrict our focus to the objects \{4,5,6\}. Figure 1(c) shows the browsing structure that could be provided at the GUI layer and Figure 1(d) sketches user interaction.

![Fig. 1. Dynamic Taxonomies](image)

The user explores or navigates the information space by setting and changing his focus. The notion of focus can be intensional or extensional. Specifically, any set of terms, i.e. any conjunction of terms (or any boolean expression of terms) is a possible focus. For example, the initial focus can be the empty, or the top term of a facet. However, the user can also start from an arbitrary set of objects, and this is the common case in the context of a WSE and our primary scenario. In that case we can say that the focus is defined extensionally. Specifically, if \( A \) is...
the result of a free text query $q$, then the interaction is based on the restriction of the faceted taxonomy on $A$ (Figure 1(b) shows the restriction of a taxonomy on the objects \{4,5,6\}). At any point during the interaction, we compute and provide to the user the immediate zoom-in/out/side points along with count information (as shown in Figure 1(d)). When the user selects one of these points then the selected term is added to the focus, and so on. Note that the user can exploit the faceted structure and at each step may decide to select a zoom point from different facet (e.g. filetype, modification date, language, web domain, etc).

Examples of applications of faceted metadata-search include: e-commerce (e.g. ebay), library and bibliographic portals (e.g. DBLP), museum portals (e.g. [12] and Europeana$^4$), mobile phone browsers (e.g. [15]), specialized search engines and portals (e.g. [20]), Semantic Web (e.g. [11,19]), general purpose WSEs (e.g. Google Base), and other frameworks (e.g. mSpace[25]).

**Related Work** Systems like [32,11,17,19,2] support multiple facets, each associated with a taxonomy which can be predefined. Moreover, the systems described in [32,11,19] support ways to configure the taxonomies that are available during browsing based on the contents of the results. Specifically, [32] enriches the values of the object descriptions with more broad terms by exploiting WordNet, [11] supports rules for deciding which facets should be used according to the type of data objects based on RDF, and [19] supports reclassification of the objects to predefined types. Furthermore, there are works [1,18] in the literature that compare automatic results clustering with guided exploration (through dynamic faceted taxonomies). However, none of these systems apply content-based results clustering, re-constructing the cluster tree taxonomy while the user explores the answer set. Instead they construct it once per each submitted query.

3 Vision: Advanced Exploratory Search

3.1 Information Space (Advancements)

**Integration of Mined and Explicit Metadata** To the best of our knowledge, there are no other WSEs that offer the same kind of information/interaction. A somehow related interaction paradigm that involves clustering is Scatter/Gather [5,10], which allows the users to select clusters, subsequently the documents of the selected clusters are clustered again, the new clusters are presented, and so on. However, the application of results clustering on thousands of snippets would have the following shortcomings: (a) **Inefficiency**, since real-time results clustering is feasible only for hundreds of snippets, and (b) **Low cluster label quality**, since the resulting labels would be too general. On the other hand, dynamic taxonomies can load and handle thousands of objects very fast [27]. To this end we propose a **dynamic (on-demand) integration** approach. The idea is to apply the result clustering algorithm only on the top-$C$ (usually $C = 100$)

---

$^4$ http://www.europeana.eu
snippets of the current focus. This approach not only can be performed fast, but it is expected to return more informative cluster labels.

The approach is described at [21], where the contribution of our work lies in: (a) proposing and motivating the need for exploiting both explicit and mined metadata during Web searching, (b) showing how automatic results clustering can be combined with the interaction paradigm of dynamic taxonomies, by clustering on-demand the top elements of the user focus, (c) providing incremental evaluation algorithms, and (d) reporting experimental results that prove the feasibility and the effectiveness of the approach.

3.2 User Interaction (Advancements)

Ranking and Reduction of Facets and Zoom Points Since our intension is to provide a rich variety of facets to the user, the ranking of the available facets and zoom points will play an important role for a positive user experience. Only the top − K facets and zoom points will be presented to the user, while options to visit any facet and zoom points will be provided. As a result, we have to provide appropriate facet and zoom points ranking methods. These methods can be based on the count information of indexed objects, sibling and children information of zoom points, lexicographic ranking of terms, or using TF*IDF weighting. Facets and zoom points reduction will be especially important for mobile devices, with limited displays.

In addition, the user will be able to define a specific threshold (i.e. by using a bar), to restrict the number of results. This threshold can be defined either for all the returned objects, or to objects of specific facets. Moreover, this threshold can be defined for specific attributes of a facet, like number of instant children, count number, etc. Using this interaction model the user could possibly get a better understanding of the returned objects and their relationships.

3.3 Interactive Preference Management

IR visualization systems should provide efficient and effective access to exploratory information needs. Ways that can extend the user actions in order to further ease the interaction, and restrict the infobase, to those parts of the information that the user is interested in, should be investigated. Currently, preference management and adaptation, requires from the user to formulate complex expressions or interact with complex UIs. The above observations justify the need for flexible and universal access methods that offer on-line preference elicitation and support. Requirements of such explorative environments include: a) simplicity (the users should be able to use and understand the interaction immediately), b) expressiveness (it should be possible for the user to interactively specify complex preference structures), and c) acceptance by users (the resulting interaction should be effective and desired by the users).

We will investigate how we can extend the user actions in order to further ease the interaction and to speed up the restriction of the focus to those parts of the information space that the user is interested in. These actions will affect
the appearance and presentation order of facets, terms (zoom-in/side points) and objects of the focus. The semantics will be described using the framework described in [16], extending it to also support hierarchically organized values. This could possibly be done by supporting a form of preference inheritance.

3.4 Exploratory Search and Visualization

The diversity of IR visualization models, poses the question if they can be synthesized into one visualization environment. Two basic strategies in this direction are identified. The first is to display multiple visual configurations simultaneously in a larger visualization environment. The second is more complex, since it synthesizes various visualization approaches into one new visualization approach. In this case, their data structures should be compatible and their displayed attributes should be complementary [35]. Recently, there is an effort to investigate the motivation and the feasibility for designing a declarative language for the specification of visualization and interaction methods which will allow the formal expression of structure, appearance, behavior and communication between the various structures of information visualization [3, 8].

We will focus in investigating ways to integrate the different visualization models into the interaction paradigm of dynamic and faceted taxonomies. In such a way, we will provide a synthesized visualization environment, which will take advantage of each visualization model’s strengths, and at the same time overcome it’s weaknesses. Furthermore, we will try to provide graph functionality through an intuitive UI, for the visual representation of 2 or more facets. The user will be able to create different types of 2D or 3D graphs, by dragging facets or zoom points to a well defined area. Moreover, he will be able to define each aspect of the graph, like type of graph, axis type, values of axis, order, etc. Such graphs can help users get an instant overview of the infobase. For example using the zoom point of an author’s name in the Author facet and the Year facet, a user could see the number of papers for a specific author, during a defined time period. For all the above, we will take into consideration psychology theories, like pre-attentive processing and Gestalt theory, for more intuitive user interaction.

3.5 Metrics for Exploratory Search

Evaluation of exploratory systems is crucial to their success and refers to measuring the extent to which people use them to achieve goals in terms, of effectiveness, efficiency and satisfaction. [14] discusses various methods and measures of controlling the experimental studies of web search interfaces and [30] proposes an evaluation method for exploratory search features. The evaluation of such systems is difficult, because of the complexity of data relationships, diversity of displayed data, interactive nature of exploratory search, along with the perceptual and cognitive abilities offered. They rely heavily on users’ ability to identify and act on exploration opportunities [29]. Important parts of retrieval results, trends, patterns, clusters, and other aggregate information, are difficult to be measured and no specific metrics are available. Finally, it is difficult to come up with
an universal evaluation system. So, the selection and definition of appropriate metrics for the evaluation of exploratory search and IR visualization systems is an important factor for the design and evaluation of exploratory services. These metrics can also facilitate appropriate automatic adaptation functionality, including the requirements of mobile devices and phones with small displays.

4 Implementation and Experimental Evaluation

The implementation will be done in the context of Mitos\(^5\) [22], which is a prototype WSE\(^6\). FleXplorer is used by Mitos for offering general purpose browsing and exploration services. Currently, and on the basis of the top-\(K\) answer of each submitted query, the following five facets are created and offered to users:

- the hierarchy or clusters derived by HSTC (one of the two STC variants we have developed)
- web domain, a hierarchy is defined (e.g. \texttt{csd.uoc.gr < uoc.gr < gr}),
- format type (e.g. pdf, html, doc, etc), no hierarchy is created in this case
- language of a document based on the encoding of a web page and
- (modification) date hierarchy.

4.1 Preliminary Experimental Results

Loading times of FleXplorer have been thoroughly measured in [27]. In brief, the computation of zoom-in points with count information is more expensive than without. In 1 sec we can compute the zoom-in points of 240,000 results with count information, while without count information we can compute the zoom-in points of 540,000 results.

Regarding the execution times that correspond to the integration of FleXplorer and results clustering using the non-incremental and an incremental approach of HSTC, for the top \(- C\) elements, they are described in [21]. It is evident that for top-100 and top-200 values, the results are presented to the user almost instantly (around 1 second), making the proposed on demand clustering method suitable as an online task. Moreover, we can see that there is a linear correlation between time cost and the top-\(C\) value. Finally, calculating and loading clusters for the top-500 documents, costs around 3 seconds making even big top-\(C\) configurations a feasible configuration. A significant speedup is observed when the incremental algorithm is used and the overlap of snippets is more than 50%.

5 Epilogue

Apart from the methodological issues for achieving the above, this dissertation will focus on (a) performance, so that the resulting methods and techniques are

\(^5\) \url{http://groogle.csd.uoc.gr:8080/mitos/}

\(^6\) Under development by the Department of Computer Science of the University of Crete and FORTH-ICS
applicable on large volumes of information, (b) flexibility of applicability, so that they are also applicable to different types of information (from simple text and unstructured data, to semi-structured and structured data), and (c) adaptability (or on-site configuration) of these services. With regard to adaptability, the operation should be simple and systematic so that it can support the adaptation of services based on the environment (context) of the user.

References

13. J. Jaanruang and W. Kreesuradej. A new web search result clustering based on true common phrase label discovery. In Procs of the Intern. Conf. on Computational Intelligence for Modelling Control and Automation and Intern. Conf. on Intelligent


