Multimed Tools Appl (2013) 62:507–543
DOI 10.1007/s11042-011-0910-2

Multiple perspective interactive search: a paradigm for exploratory search and information retrieval on the web

Rahul Singh · Ya-Wen Hsu · Naureen Moon

Published online: 10 November 2011

Abstract The World Wide Web (WWW) represents the largest and arguably the most complex repository of content at our current state of technological development. Information on the web is represented using a variety of media, with a (current) predominance of text- and images-based data and increasing presence of other media such as video and audio. The complexity and heterogeneity of the information implies that the associated semantics is often user-dependent and emergent. Thus, there is a need to develop novel paradigms for web-based user-data interactions that emphasize user context and interactivity with the goal of facilitating exploration, interpretation, retrieval, and assimilation of information. This article presents a novel presentation-interaction paradigm for exploratory web search which allows simultaneous and semantically correlated presentation of query results from different semantic perspectives. Users can explore the results either using a specific perspective or through a combination of perspectives via highly-intuitive yet powerful interaction operators. In the proposed paradigm, hits obtained from executing a query are first analyzed to determine latent content-based correlations between the pages. Next, the pages are analyzed to extract different types of perceptual and informational cues. This information is used to organize and present the results through an interactive and reflective user interface which supports both exploration and search. Experimental investigations, many of which are conducted in comparative settings, analyze the proposed approach in query-retrieval scenarios involving complex information goals. These results demonstrate the efficacy of the proposed approach and provide important insights for the development of the next-generation of interfaces for web-search.

Keywords Exploratory web search · Human-computer interaction · Information retrieval · Experiential interfaces · Direct manipulation · Multiple perspective search · Information visualization · Spatial search · Temporal search · Multimedia information systems · Information goal · User-media interaction

R. Singh (✉) · Y.-W. Hsu · N. Moon
Department of Computer Science, San Francisco State University, San Francisco, CA 94132, USA
e-mail: rahul@sfsu.edu

1 Introduction

Users seek information on the web through two predominant modes [34]: by browsing, called "search by-navigation" or by searching, called "search by-query". In the first mode, the interaction between the user and the data repository is driven directly by the user's interpretation of their information need and their information foraging constraints. In the latter mode, a search engine typically mediates the user-data interactions and the process starts with the user entering query-terms that act as surrogates for the user information goals. Given a query, the most common strategy has been to present the results as a list where each entry is ranked by its putative relevance to the query. Users have to subsequently peruse the list to satisfy their information needs through browsing the links and/or by issuing further queries.

Such a user-data interaction paradigm is perfectly adequate for many types of queries. However, as the information in the web gets diversified both in terms of its complexity as well as in terms of the media through which the information is encoded, short keyword-based queries are often insufficient to describe both the user information need as well as the content that may putatively satisfy it. It is also interesting to note here that even under conditions where users have a well defined information goal, many may not necessarily employ keywords as the first-choice modality. This observation was made in [46], where in a modified diary study that examined how people performed personally motivated searches, it was found that instead of jumping directly to their information using keywords, a majority of participants navigated (to the target) using small local steps using their contextual knowledge. In recognition of these issues, commercial search engines have begun to provide alternatives; for instance, augmenting the listing of results with page thumbnails or previews.

The user information need can, of course, be more nuanced than one that simply requires looking-up clearly defined facts. In such cases, the user may choose to explore the information space and the information goal may itself evolve as part of the exploration process. Finally, and most fundamentally, the semantics that can be associated with media and even complex alphanumeric data can rarely ever be predefined and kept fixed. As has been demonstrated [9, 37], the semantics of such data are non-unique, user-dependent, and *emergent*. Emergent semantics implies that data is endowed with meaning by placing it in context of other similar data and through factors that are user specific.

Given this context, the limitations of the strategy of simply presenting a list of results, even when accompanied with cues, becomes apparent. From an operational perspective alone, such a paradigm increases the cognitive load on users by forcing them to cherry-pick from a list that may include a variety of hits not all of which may be related to their information goal. The problem is exacerbated if the query terms are polysemous or if the results contain multiple topics. More fundamentally, the richness of the information-seeking task in a web containing heterogeneous and multifarious information requires that user context and user-data interactions play a critical role in interpretation and assimilation of the information. A similar conclusion can also be reached if one examines the issue from the perspective of explanatory modeling of user behavior on the web. For example, research in *information foraging* theory [28] asserts that users typically navigate towards their information goal by following *link cues*, which are fragments of information within or near hyperlinks and relevant to the user's information goal. From this perspective also, the conventional approach arguably remains minimalistic in that it neither presents any clues about

correlations within the list of results nor does it provide sufficiently rich link cues. Two distinct but interrelated necessities can thus be identified:

- Capturing and representing the variability in the semantics of the information that may be spread across different web pages which have been identified to be of relevance to the query.
- Supporting efficient and effective interactions between users and the information with the ultimate goal of efficiently satisfying the user information need even if it is not well defined at the beginning of the search process.

Both the aforementioned issues underline the need for facilitating exploratory, user-centric capabilities rather than pure syntactic query-retrieval. In [22], Marchionini had proposed three types of search activities: *lookup*, involving carefully specified queries and precise results, *learning*, involving cognitive processing and knowledge interpretation, and *investigation*, requiring critical assessment. It was postulated in [22] that information seeking activity involving learning and investigation constitutes exploratory search. It has also been noted that in exploratory search, information seeking is not just about the final results but also about knowledge acquisition during the search process itself [50, 51].

The development of novel paradigms in this context, arguably, has to build on the available search technologies. This would allow taking advantage not only of the well developed keyword-based information retrieval capabilities of modern search engines but also of their capabilities to crawl and index the web. Furthermore, our current inability to bridge the signal-to-symbol gap [43], implies that highly efficacious media-content-based approaches for querying the media-rich web are, as of yet, infeasible in general settings.

In this paper, we build upon our prior work [14, 38, 39] and present a paradigm for user-data interaction in web-search which is motivated by the philosophy of experiential interfaces proposed for media-rich environments [15, 40]. In the proposed approach, we empower users to search and interact with web-based information from multiple semantically relevant perspectives. These perspectives can be data driven, model-based, or a combination of the two. Furthermore, two of the perspectives supported in this paradigm are dedicated to the support of spatial and temporal reasoning with the data, since, space and time are central to perceptual reasoning. Finally, the proposed paradigm emphasizes interactivity; outcomes of interactions conducted with respect to a chosen semantic perspective (such as, for instance, the temporal characteristics of the data) are reflected in the other semantic perspectives. Thus, users can experience the underlying relationships between different aspects of the information and form a holistic understanding of the information. Experiments and user studies indicate that such an approach can be especially helpful in situations such as exploratory search, finding media-based information, understanding the information space induced by the query, and recognizing the information diversity underlying the query. For complex information needs, the proposed paradigm also appears to facilitate more rapid identification and retrieval of relevant information when compared to the classical approach of ranking and listing of the results.

2 Review of prior research

2.1 Visualization and retrieval of text and media

The problem of presenting complex information in manners that aid assimilation and retrieval has been actively investigated in the context of text and media visualization for digital libraries. Examples of early ideas in text visualization include the starfield displays

[1], TileBars [11], use of the Kohonen self-organizing map to project the high-dimensional document space to a 2D plane for visualization [21], and the SPIRE Galaxies visualization which mapped document time to line angle and length [53]. Unlike these methods, which focused on the visualization of specific document characteristics, the Envision system [25], supported representation of user-selected characteristics. Other techniques have focused on capturing more complex and abstract document characteristics. For instance, the ThemeRiver visualization approach was proposed to analyze large document collections and depict thematic variations over time so as to help users find trends and cause-effect relationships [10]. However, many of these tools were not designed to factor the user and data contexts, the history of interactions, or the evolving nature of the user information goal.

Given the rapid increase in media-based information on the web, a number of strategies have also been proposed to address the problem of web-based media retrieval. For instance, in the AMORE search engine [24], different strategies were explored for assigning keywords to images to facilitate retrieval. A combination of color-texture moments, text, and link information was proposed to cluster image search results from the web in [3]. Another content-based approach was proposed in [49], where the images were segmented into homogeneous regions and quantized into codewords. The collection of such codewords was then ranked based on human labeled data and used to cluster other images. The use of context to retrieve media provides an alternative to the content-based approaches. Examples include the use of location information for image retrieval [47], the use of temporal information [8, 29] for indexing and retrieval of media, and the use of event-based unified indexing of multiple semantically correlated media [41]. We refer the interested reader to the reviews [20, 44] for further details on media retrieval. As the focus of this research is the design of methods for visualization and interaction with web-search results, the following subsection reviews the relevant research.

2.2 Clustering and visualization of web search results

The fact that alternate presentation of search results (as opposed to the use of a linear list) tends to support exploration has been pointed out in many studies. In one of the early investigations [32], a tabular interface was compared with a conventional list-based presentation of results. The columns of the table in this case corresponded to elements of the search results such as title, URL, metadata, excerpts etc. It was observed that the tabular interface induced users to undertake more diverse search strategies than the list-based presentation. More recently, the list format was compared to a grid-based presentation to study the role of the interface in the user's evaluation process [16]. In this study the trustworthiness order and search interface was varied, leading to four between-subject factors: presentation in terms of descending trustworthiness, presentation in terms of ascending trustworthiness, list-based interface, and grid-based interface. In the list interface, users were found to give the greatest attention to results placed at the top of the list. Consequently, in the case of presentation in terms of ascending trustworthiness, users spent significantly longer time on the least trustworthy results. In contrast, with the grid interface all results were attended to equally long. Thus, in general, the grid interface was postulated to better support users in the selection of trustworthy pages. A similar conclusion was also reached in [35] where the Kartoo (www.kartoo.com) interface was compared with a Google-like interface. In the Kartoo interface, the results are presented as a constellation of documents, in which the existing relations between pages are made explicit. Undergraduate students tasked with reading a set of web pages on climate change were found to

comprehend the material better using a Kartoo-like presentation as opposed to a Google-like presentation.

In the past decade, there has been an increasing interest in exploratory web search. Introductions to this area can be found in [50–52]. The demands of supporting exploratory search have spurred research on analysis, visualization, and interaction techniques for web-based information retrieval. In the following, we consider these works in terms of two, sometimes overlapping, directions. The first of these includes methods that use text analysis to cluster search results based on the page content. Clustering of results not only helps users form an overview but has also been found to improve the speed of retrieval [6]. The second direction involves methods that focus primarily on visualization of search results. In addition to the techniques reviewed in this section, the interested reader is referred to [4] for further details on web page clustering.

The traditional approach to document clustering involves manual classification using a taxonomy, or hierarchy of descriptors. With regard to web pages, this approach although semantically very meaningful, is limiting because developing taxonomies and assigning them to web pages is very costly. Moreover, manual classification cannot keep pace with the rapidly increasing number of documents on the web. In the Open Directory Project (www.dmoz.org), these limitations are ameliorated through taxonomy-based classification using a large global community of volunteer editors. In spite of limitations, the utility of ODP classifications is significant as underscored by its use in Google Directory, AOL Search, as well as in the proposed method.

Algorithmically clustering search results so that each cluster putatively corresponds to a distinct topic is an alternate strategy that does not suffer from limitations inherent to manual page classification. An example of this approach is the Grouper interface [56], where search results are clustered and the clusters are labeled using phrases extracted from page snippets. Other efforts include [7], where a hierarchical clustering of results using web page summaries was used. Since the quality of clustering ultimately depends on page descriptors and clustering strategies, significant research has occurred in these two directions. In [31], position sensitive word-based and TFIDF-based descriptors were proposed for clustering. Instead of using word-level descriptors, the clustering problem was formulated as that of ranking salient phrases in [57], while in [17] web pages were classified using their URLs. An approach towards grouping search results based on a cognitive model of language was proposed in [18], where the search results were filtered using WordNet (http://wordnet.princeton.edu) senses. Unlike these approaches, where the clustering and categorization was defined by page content, in the search engine Northern Light (www.northernlight.com), the results were clustered into categories predefined in library sciences. A number of commercial efforts have also supported on the-fly clustering of search results. Examples include Clusty (www.clusty.com), Vivisimo (www.vivisimo.com) and Grokker (www.grokker.com).

The importance of visualization in web search was highlighted in the early parts of the last decade, when in [54] Woodruff et al. showed that engaging the perceptual capabilities of users by using thumbnails of web-pages aided them in finding relevant information after a search engine had produced a list of hits. Currently search engines, such as ASK (www.ask.com) or Google include thumbnails of pages retrieved by a query. A number of research efforts have also investigated the use of more powerful visualization techniques. For example, in [12] a variation of heat-maps was used to represent web-search results and support their interactive exploration. In [19], the authors proposed the Venn Diagram Interface (VDI). In this clickable interface, the number of hits generated by terms in the query was shown in each subset of the diagram. The idea behind VDI was to allow users to see how terms in the query contributed to the result set

of the query and subsequently help them browse the results corresponding to each of the query terms. The WordBars system [13] was also designed to present the distribution of results in relation to the terms of the corresponding query and facilitate browsing of the results. In this system, users were presented with a list or bar-graph of the most frequently occurring terms in the result set of a given query. Users could then sort the list using one or more of the terms as an index or reformulate their query by adding one or more terms from the list or bar to their original query. Researchers have also used abstract graphical object (glyph) representations to map the value of each attribute of an input tuple to a visual dimension [5, 33]. For instance in [5], a flower metaphor was used for the glyph design and the characteristics and metadata of the web documents (such as document length and type, domain, frequency of keywords from the query, and number of links) constituted the attributes being mapped. Unlike the above methods, which were all designed for generic search, in [26], a specific type of exploratory search was considered which occurs when, for example, a user searches the web with the goal of collecting information on different cars. The authors postulated that the cognitive workload in such cases is primarily due to factors such as the need to represent information goals, determine how informative a specific page is, and memorize previously explored information. To mitigate the workload in such cases, a system called SketchBrain was proposed. In this system the query trails and post-query browsing trails were tracked, stored, and visualized as a sketch. The system used a path-recommendation algorithm to help users choose the next page to visit. The users could also interact with and manipulate the abstracted entities in SketchBrain using operations such as projection, selection, and associations with user defined or system recommended topics. At the state-of-the-art, a number of commercial search engines have begun to support visualization of search results to varying degrees. On one end systems such as the aforementioned Vivisimo, Grokker, and Kartoo have experimented with incorporating significant visualizations strategies in their interfaces. On the other end, mainstream search engines like Bing and Google have started to combine listings with perceptually relevant presentations such as images and maps. Finally, we would like to note faceted browsing [55] which is an information presentation technique where the attributes corresponding to a specific entity is organized along multiple orthogonal dimensions, called facets. The collection can be browsed by selecting values in the facets. The advantage of faceted browsing is that given a query, the available values for a corresponding constraint are always communicated to the user. The display of information in terms of information facets was shown in [55] to be more efficacious as compared to a keyword search interface. However, faceted classifications tend to be limiting when the collection is large or when users have varying information needs.

3 Design principles and overview of the prototype system

3.1 Experiential computing and the design principles

Supporting the information seeking process, once it moves out of the realm of queries that can be precisely specified and enters the exploratory domain requires facilitating exploratory and user-centric capabilities. In multimedia computing, the principles of *experiential computing* were proposed to address the challenges of information exploration, assimilation, and retrieval in settings involving heterogeneous and multifarious data [15, 40, 42]. This paradigm argues for the design of systems where users can apply their natural

senses to observe, interact, and explore the data. Our approach is motivated by this paradigm.

Experiential systems are characterized by the following properties: (1) they are direct, in that they do not use complex metaphors or commands either for presentation of the information or for mediating interactions with it, (2) they support the same query and presentation spaces to support intuitive and direct user-data interactions, (3) they maintain user state and context, (4) they present information independent of (but not excluding) media type and data sources, (5) they provide multiple semantic perspectives on the data, both for presentation and interactions, and (6) they seamlessly integrate powerful algorithmic analysis with visualization and interaction.

For our problem context, experiential computing shares many characteristics with ideas proposed in visualization research. Notwithstanding its origins, it should consequently be thought of as a paradigm that encapsulates information visualization and exploration. Its relative uniqueness lies in the emphasis it places on information presentation in manners that are both perceptual and avoid the use of complex metaphors. The rationale underlying this emphasis is that the cognitive mechanism is well acclimatized to deal with direct information presentation and interaction, both evolutionarily as well as in terms of life experiences. This design principle is supported by the cognitive fit theory [48], which suggests that users achieve better task performance when they do not need to transform the model through which information is presented to a different mental model, in order to solve a task. Another key characteristic of experiential computing lies in facilitating human-machine synergy by combining powerful algorithmic data processing and analysis with interfaces that allow users to leverage their perceptual abilities for exploration and assimilation. Recently, a similar argument was (independently) made by Perer and Shneiderman to incorporate statistical computing in visualization for exploratory data analysis [27].

3.2 Overview of the prototype system

In the following we provide an overview of how an experiential user-data interaction paradigm was developed as part of this research and implemented in a prototype system. The key modules of the system and the information flow between them are shown in Fig. 1. A snapshot of the user interface is shown in Fig. 2. Queries issued to the system are directed to a search engine, selected by the user, with the *Web Data Retrieval* module obtaining the content of the pages constituting the hits. This module also uses an external service to obtain web page thumbnails. The information from the retrieved pages is next processed successively by the *Textual Information Analysis*, *Location Extraction and Analysis*, and *Time Extraction* modules. Briefly, the Text Analysis module analyzes the retrieved results to determine semantic correlations between them. The Location Extraction module parses each document and cross-references the terms with a comprehensive gazetteer of locations to find geographical associations. Similarly, the Time Extraction module parses each document to extract any dates. The design of these modules and the underlying algorithms are described in detail in Section 4.

The information derived from the various components is displayed as separate panels in the user interface (see Fig. 2), with search results presented as clusters of web pages, and related spatial and temporal information shown as points on a map and timeline. As users mouse-over links, a fourth panel presents a thumbnail of the web-page, a summary of its content, and any media files (such as audio or video) that may be associated with it. The design of the interface has been done to support emergent and exploratory user interactions.

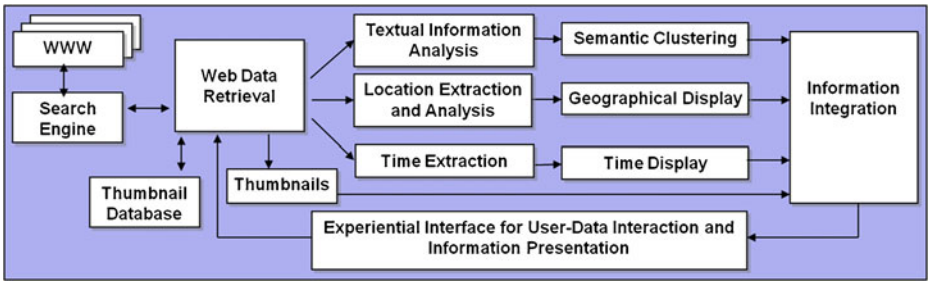


Fig. 1 The key components of the system and the information flow between them. A user issues a query and interacts with the results using the experiential user interface shown in Fig. 2. The query is passed on to a search engine by the web data retrieval module. This module also collects the returned results and interfaces with a thumbnail database to retrieve the thumbnails of the pages returned by the search engine. Subsequently, the retrieved pages are analyzed in terms of their content and any location and time information that may be contained in them. The content of the pages is used to semantically cluster them. These clusters, along with the geographical display, time-information display, and thumbnails are brought together by the information integration module and presented to the user for subsequent interactions

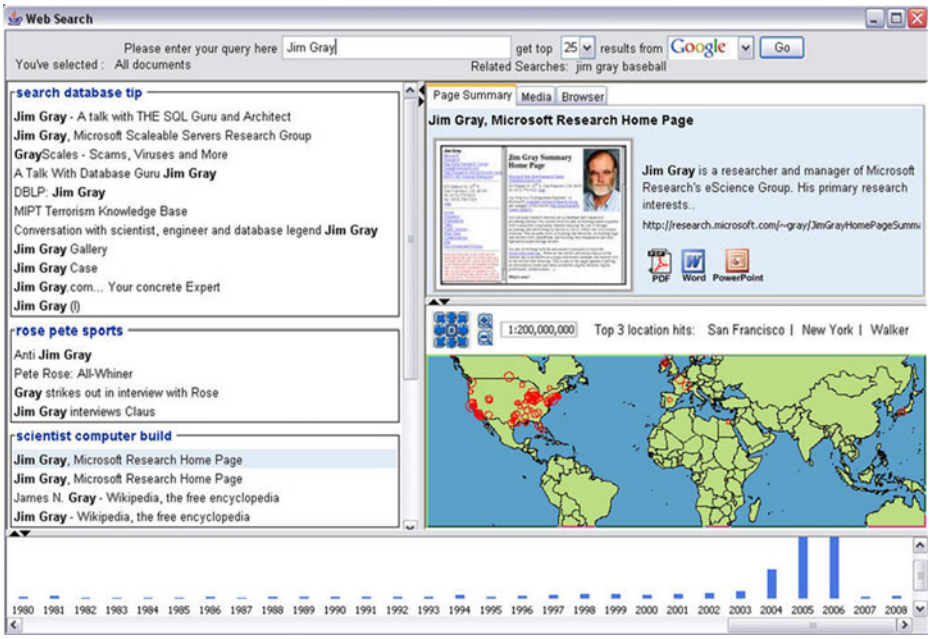


Fig. 2 The experiential user interface of the proposed prototype showing the different coordinated views of the search results. The top-left panel of the interface displays the semantic clustering of the search results corresponding to the user query (in this case "Jim Gray"). As users mouse over any specific URL in any of the clusters, a page summary is created. The summary includes links to any media present in the page and is displayed in the top-right panel. The geographical distribution of information from the selected cluster of hits is presented through an interactive map. Finally, the temporal distribution of information from the selected cluster is shown in the bottom panel. Users can directly interact with the information in each of these panels. Changes brought about as a result of interactions in any one of the panels are reflected in the other panels. For instance, selection of a specific cluster updates the timeline as well as the map to display the temporal and geographical information related to this cluster

In it, various views of the data are tightly linked to each other, so that interactions in terms of any one of them are instantaneously reflected in all the views. For example, selecting a specific region of time leads all links relevant to it to be highlighted, including in the spatial view. Such coupling helps users perceive the correlations between different aspects of the information. Consequently, it can help in formulation of hypotheses and in the discovery of relationships in the data.

4 Determining the semantics of the information structure underlying the query

In order to determine the semantic structure of the information retrieved by executing the query, it is essential to determine web-pages that have similar content and distinguish them from those that contain semantically different content. Typically the heterogeneity in the semantic structure underlying a query is due to polysemous terms or, more fundamentally, due to the inherent variability of the information corresponding to the query terms.

As a preprocessing step towards helping users visualize and interact with the information structure underlying a query, the text of each retrieved web page is appended to its metadata and parsed to remove all html tags, numbers, and punctuation. A stop-list of 670 common English words is next used to remove minimally-descriptive terms (e.g., "that"). Of the remaining words in the page, the first 200 words are used for subsequent processing. Next, the Porter stemming algorithm [30], is used to truncate words to their root forms in order to allow morphological variants to be mapped together for frequency counting. The modified keywords are used to construct a term-frequency table where each column denotes a document and each row a keyword. Following the preprocessing step, a novel term frequency adjustment scheme is employed. This step combines Latent Semantic Indexing (LSA) with a Term Frequency-Inverse Document Frequency (TFIDF) weighting to enhance the quality of the final document clusters by increasing the similarity between the term frequency vectors of similar documents and simultaneously amplifying the differences between the term frequency vectors of dissimilar documents. The intuition behind the scheme lies in utilizing LSA to first reveal semantic correlations between the documents via mapping to a low-dimensional Eigenspace. Next, the prominence of uncommon words is further emphasized through TFIDF. The following subsections describe this method in detail.

4.1 Determining term saliency by TFIDF weighting in the latent semantic subspace

After normalization (dividing frequency values by document size), LSA is performed on the term frequency matrix $\boldsymbol{X}$ (with m rows (terms) and n columns (pages)). The technique consists of performing singular value decomposition (SVD) on the matrix $\boldsymbol{X}$ to rewrite it as a product of 3 matrices as shown in Eq. (1).

$$X = U \times \Sigma \times V^T \tag{1}$$

In Eq. (1), $\boldsymbol{U}$ is a unitary matrix of size $m \times m$, $\boldsymbol{\Sigma}$ denotes the diagonal matrix of scaling values of size $m \times n$, and $\boldsymbol{V}^T$ denotes the conjugate transpose of $\boldsymbol{V}$ and is of size $n \times n$. The dimensionality of the data is reduced by truncating all but the largest eigenvectors, so as to minimize the amount of noise in the data set. We automatically select the reduced number of dimensions by calculating the difference between contiguous eigenvalues and truncating all values after the largest drop. Thus, the dimensionality reduction process is completely

data-driven. Finally, the matrix is reconstructed to reveal the modified term frequency values. In the next step of our analysis, the term prominence values are adjusted using TFIDF [36]. Thus, our content analysis approach combines LSA and TFIDF. In the following, we briefly describe the TFIDF method and explore specific questions related to its application to our problem domain and then illustrate the process of combining LSA and TFIDF using an example.

The fundamental idea underlying TFIDF is based on the observation that uncommon terms give better discernment between documents. This method uses a background set of documents to down-weight common terms. The background set is constructed from the same domain as the test data, and term frequencies are adjusted as shown in Eq. (2):

$$TFIDF(i,j) = TF(i,j) \times \log_{10}\left(\frac{N}{DF(i)}\right) \tag{2}$$

In Eq. (2), *TF(i,j)* denotes the frequency of term *i* in page *j, N* is the size of background set, and *DF(i)* denotes the number of background set documents in which term *i* appears. The last multiplicand is called the inverse document frequency. An important characteristic of the TFIDF weighting scheme is the fact that its performance is heavily influenced by the choice of background set. In the case of information gathering on the web, no a priori domain of discourse can be fixed. Thus, the strategy for selecting the background set is crucial. We experimented with two methods to determine the background set. In the first case, the background set was created using 1656 randomly selected web pages. In the second case, the background set was dynamically constructed for each query and consisted of terms from the first 50 pages returned for the given query. For both the cases, content of the pages were pre-processed through stop-listing and removal of duplicate terms.

Our idea for dynamically determining the background set is motivated by the hypothesis that the discernment value of a term depends upon the context in which it appears. For example, the term "quark" is uncommon in normal English usage. However, this term would occur frequently in the results for a query on particle physics. So, the importance of this term for distinguishing pages in the particle physics domain arguably would be limited. In Fig. 3, we present results from 40 distinct and diverse queries where the list of hits was clustered using Person correlation of the TFIDF term weights in each page. Two sets of TFIDF weights were calculated following each of the strategies for background set creation described above. As can be seen from Fig. 3 (Top), better results were consistently obtained with the dynamically constructed background set.

We next combine LSA and TFIDF (abbreviated henceforth as LSA + TFIDF) to emphasize the difference between dissimilar pages while simultaneously increasing the similarity of related pages. Our idea is motivated by the observation that the difference between inter-cluster and intra-cluster correlation values can often be low, when LSA alone is employed. Examples include the 1st, 9th, 10th, 14th, and 35th queries, corresponding to the query terms "eclipse", "jobs in bay area", "stock car racing", "Japanese comics" and "chili pepper" as shown in Fig. 3 (bottom). The same figure also shows the improvements obtained by applying TFIDF weighting after LSA. For instance, the reader can note that the combined method led to consistently higher correlation values for similar documents while emphasizing the difference between dissimilar documents.

In the following, we illustrate the effect of LSA + TFIDF through a small real-world example, where the polysemous word "eclipse" was used as the search term. In Table 1, we present the key terms for the first five pages retrieved for this query. We used 10 terms (after stop-listing) to constitute the representation space for the documents. Their term frequency

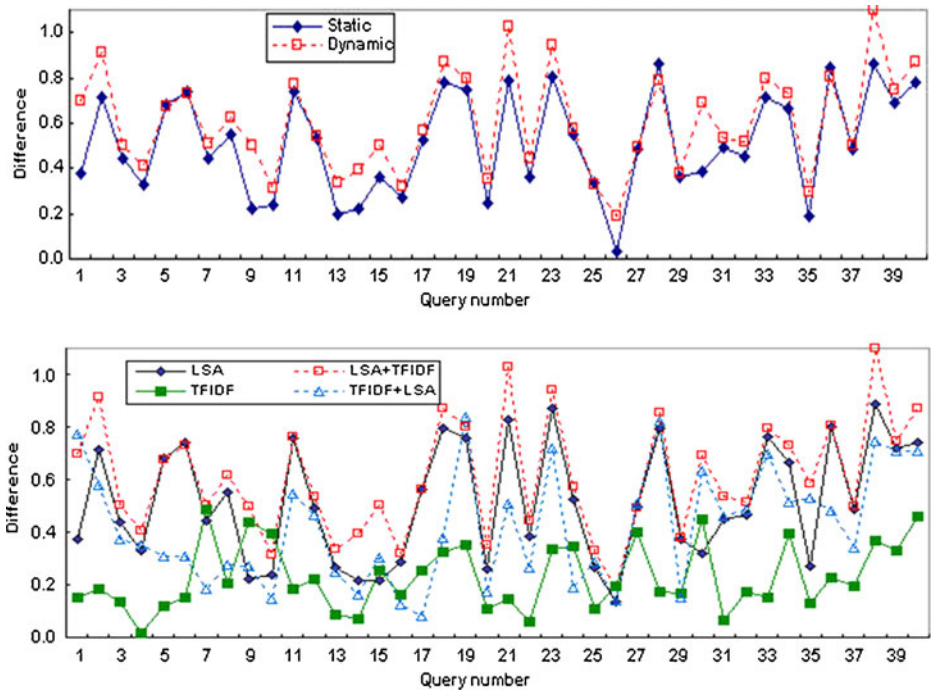


Fig. 3 (*Top*) Comparison of static and dynamically generated background sets for computing TFIDF values across 40 different and diverse queries. The static background set was created using terms from 1656 randomly selected web pages and did not change across queries. The dynamic background set was generated for each query and consisted of the terms from the first 50 search results (pages). The X-axis displays the query number and the Y-axis shows the average inter-cluster distance based on the TFIDF scores for terms in each document. Larger inter-cluster distances indicate better term-weighting. It may be noted that for all the cases, the use of a dynamic background set led to inter-cluster distances that were equal or greater than those obtained using the (larger) static background set. (*Bottom*): Comparison of LSA, TFIDF and their combinations

values are shown in Table 2. Furthermore, the first 20 terms from each of the first 10 retrieved pages were used as the background set (not shown here). It is easy to observe from Table 1 that the web pages fall into two categories: *software* (first 3 documents) and *astronomical*

Table 1 Terms after stop-listing from the first five documents of the query "Eclipse"

Page	First 20 words after stop-listing
1	news featured corner articles kind universal tool platform open extensible ide particular check roadmap white paper read technical articles visit
2	downloads downloads section start downloading sdk browse various project pages useful tools plugins need problems installing getting workbench run check
3	subclipse tigris login register collabnet enterprise edition tigris open source software engineering tools pages projects community projects subclipse project tools
4	nasa solar lunar resource planetary transit year ephemeris moon phases sunearth gsfc nasa elcome nasa gsfc sun earth connection education
5	solar stories path totality exploratorium nasa sun earth education forum presents live webcast June total solar zambia maps features resources

Table 2 The raw term frequency scores of the ten terms from Table 1 used for representing he documents

Terms	Page1	Page2	Page3	Page4	Page5
Open	1	0	1	0	0
Check	1	1	0	0	0
Project	0	1	1	0	0
Pages	0	1	1	0	0
Tools	0	1	2	0	0
Nasa	0	0	0	3	1
Solar	0	0	0	1	2
Sun	0	0	0	1	1
Earth	0	0	0	1	1
Education	0	0	0	1	1

events (document 4 and 5). In Table 3, we show the inter-page correlation distances based on the term weights after LSA. The reader can see the clear grouping of pages 1–3 and pages 4–5. In Table 4, the inter-page correlation values are recomputed, this time, based on terms weights obtained using LSA + TFIDF. The reader may observe the improvements in the correlation scores of semantically related pages even for this small example. For instance, the correlation value between page 1 and page 4 (which were unrelated) decreased from −0.79 to −0.89 while the correlation values between page 1 and pages 2–3 increased from 0.99 to 1.0. In the next section, we describe how the term saliency scores defined through LSA + TFIDF can be used to group semantically related content.

4.2 Clustering semantically related content

We present a two-stage clustering mechanism for grouping pages having semantically related content. The first stage of this approach is dynamic and data-driven while the second stage is model-based and uses a manually constructed taxonomy to refine the results. The second stage is crucial because it allows similar pages to be brought together even if there is a paucity of correlating data as extracted from them. The second stage also allows us to capture notions of semantic similarity that are cognitively obvious but are difficult to discern purely from the data without conceptual models.

4.2.1 A measure for comparing page-content

An important requirement for clustering web-pages in semantically related groups is the definition of an appropriate measure (or metric) between pages. This measure needs to be computable using the page descriptors. Furthermore, the measure should correspond to cognitive notions of semantic similarity. Towards this in the following, we present results

Table 3 The inter-page correlation distance after LSA

	Page1	Page2	Page3	Page4	Page5
Page1	1	0.99	0.99	−0.79	−0.77
Page2	0.99	1	1	−0.85	−0.84
Page3	0.99	1	1	−0.86	−0.84
Page4	−0.79	−0.85	−0.86	1	1
Page5	−0.77	−0.84	−0.84	1	1

Table 4 The inter-page correlation distance after LSA + TFIDF

	Page1	Page2	Page3	Page4	Page5
Page1	1	1	1	−0.89	−0.87
Page2	1	1	1	−0.92	−0.90
Page3	1	1	1	−0.92	−0.91
Page4	−0.89	−0.92	−0.92	1	1
Page5	−0.87	−0.90	−0.91	1	1

from investigations that compared three similarity measures commonly used in the text-retrieval community: Pearson-correlation, Cosine-distance, and Euclidean-distance. Giving two vectors y_a and y_b, the Pearson correlation measures their similarity in terms of a correlation score $S^{(P)}(y_a, y_b) \in [-1, 1]$. Following [45], we normalize the Pearson correlation value to the interval [0, 1] for purposes of comparison with the cosine and Euclidean measurements. The normalization from is shown in Eq. (3), where $\bar{y}$ denotes the average feature value of vector y.

$$s^{(P)}(y_a, y_b) = \frac{1}{2}\left(\frac{(y_a - \bar{y}_a)^T (y_b - \bar{y}_b)}{\|y_a - \bar{y}_a\|_2 \cdot \|y_b - \bar{y}_b\|_2} + 1\right) \tag{3}$$

The cosine measure, denoted hereafter as $S^{(C)}(y_a, y_b)$ determines the similarity of two vectors y_a and y_b by the angle between them. We also normalize the cosine measure to the interval [0,1] as shown below in Eq.(4). In this equation, a value of 1 denotes perfect similarity while the value 0 denotes complete dissimilarity.

$$s^{(C)}(y_a, y_b) = \frac{1}{2}\left(\frac{y_a^T y_b}{\|y_a\| \cdot \|y_b\|} + 1\right) \tag{4}$$

Finally, the Euclidean metric $S^{(E)}(y_a, y_b)$ measures the distance between vectors y_a and y_b. Consequently, it is unbounded on the positive side and perfect similarity is given by a distance of 0. In order to compare it with the aforementioned measures, following [45], we map the Euclidean distance to a [0, 1] interval as shown in Eq. (5).

$$s^{(E)}(y_a, y_b) = \frac{1}{1 + \|y_a - y_b\|_2} \tag{5}$$

In order to compare these measures, we used a set of 40 queries on real-world topics and manually clustered the first 20 hits for each query to obtain the ground-truth clusters. In order to mirror typical web search usage, the test queries were constrained to consist of 1 to 4 words. Examples of queries included, among others: “chili pepper”, “Japanese comics”, “eclipse”, “computer”, “cell phone”, “earthquake”, “world cup”, “jobs in bay area”, “stock car racing”, “New York”, “sports car”, “ant”, “brother”, and “wolf”. The manual clustering was done by examining the content of the hits and grouping together pages with similar content. Since cluster definitions can be subjective, we used a committee-voting approach to define the ground truth. The committee was comprised of the three authors, who independently clustered the data. The final clusters used in the analysis were obtained using majority voting. This process is further explained in Section 6.1.

For each of the aforementioned distance measures, we calculated the average correlation of each page with pages in the same ground-truth cluster as well as the average correlation with pages outside the cluster. We defined the best measure to be the one that maximizes

the difference between the within-group and between-group average correlation values. The results from this experiment are presented in Fig. 4. For the Pearson correlation distance, the average difference and variance for across all the queries was .27±0.02. The corresponding values for the cosine distance and the Euclidean distance were 0.20±0.01 and 0.06±0.004 respectively. In these experiments the Pearson measure performed better than both the cosine distance and the Euclidean distance for all the cases. We therefore used the Pearson correlation as the inter-page distance.

4.2.2 Content-driven clustering to determine semantically related pages

To implement content-driven clustering of the pages returned as the result of a query, we propose an adaptive K-medoid clustering algorithm. This algorithm is designed to address two major issues associated with the *K*-means family of algorithms, namely, specifying in advance *K,* the number of clusters and ameliorating the dependence of the final result on the cluster initialization.

We motivate the first issue by noting that different queries would result in different numbers of retrieved documents, which in turn would induce a different value of *K* for each situation. Additionally, an incorrect specification of *K* could degrade the quality of the clusters. In order to solve this problem, we use an adaptive, data-driven approach to determine *K*. The second issue we need to address relates to the sensitivity of the final clustering to cluster initialization. This problem becomes especially critical in our problem context since each document is represented by a high-dimensional term frequency vector. A well understood characteristic of high-dimensional spaces is that with increasing dimensionality, the interior of the space becomes sparse. Therefore, in order to avoid initializing clusters in the sparse regions, we employ a density based approach to identify the *K*-highest density points in the distribution to seed the clusters. As an additional modification, we employ the *K*-medoids algorithm, which is a variation of the *K*-means algorithm and uses actual data points as centers of clusters. The advantage of this strategy lies in the fact that it reduces the re-computation of the distances between objects and cluster medoids, since the medoids change less often than means. At each step of the algorithm, a stability criterion is used to split and merge clusters. A cluster is defined to be "stable" if for any two documents in the cluster, the similarity value is higher than a threshold *T* and the similarity value of a document in this cluster with any document in

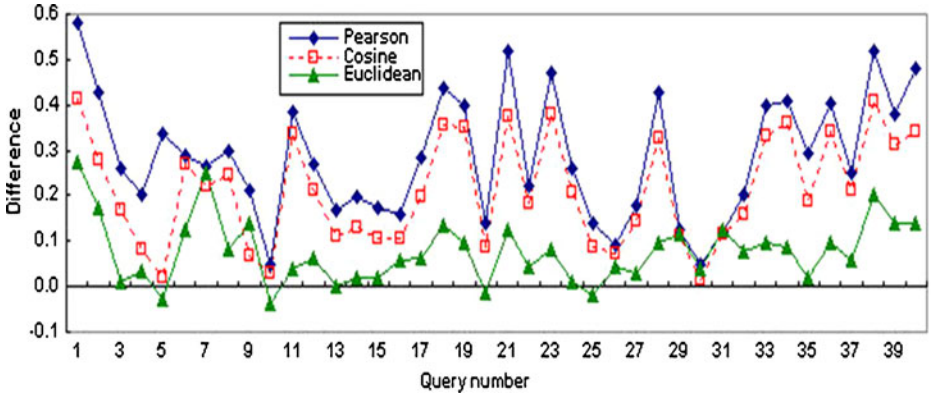


Fig. 4 Comparison of Pearson, Cosine and Euclidean similarity measures

another cluster is less than T. In the following, the set of clusters obtained after each iteration is denoted by FC. The algorithm takes as input, the data set D and the similarity threshold T. It is initialized with $K=1$ and $FC = \{D\}$. As the initial step, the data is checked for stability. If the conditions for stability do not hold (that is, there is more than 1 cluster), the value of K is incremented and the following steps are executed till all clusters become stable:

- *Initialization of cluster centroids*: The cluster centroids are seeded by selecting the K points of highest density. The density-based initialization ensures that the clustering is focused on key-regions, rather than around sparse outliers.
- *Voronoi Tessellation*: The pages are partitioned into Voronoi regions (clusters) induced by the centroids computed above.
- *Medoid computation*: For each Voronoi region (cluster) c its medoid m_c is determined as the page for which the sum of similarity values to all other pages in the cluster is the greatest:

$$m_c = \arg\max_i \sum_j S(d_i, d_j) \tag{6}$$

- *Merging and splitting of clusters*: For a given cluster, if any two objects in the cluster have similarity value lower than T, then the cluster is considered as a candidate for splitting. Analogously, for two clusters, if all pages have pair-wise similarity values higher than T, then these clusters become candidates for merging. The value of K is incremented or decremented appropriately.

The reader may note that in the above algorithm, the point at which all clusters become stable automatically determines the optimal value for K and acts as the stopping criterion.

4.2.3 Model-based grouping of pages into semantically related clusters

The variability of syntax and factors such as lack of sufficiently detailed information implies that a purely data-driven approach may not always group related pages together. To address this issue we propose a secondary grouping step based on page category information from ODP (www.dmoz.org). The ODP is a large, manually-constructed directory consisting of over 5 million web sites and maintained by a worldwide community of volunteers. The directory consists of a hierarchy of categories to which web sites are assigned. For example, the Monterey Bay Aquarium web site is classified as "*Aquariums/North America/United States/California*". Similarly, the website of the San Francisco State University is categorized as "*Reference/Education/Colleges and Universities/North America/United States/California/California State University.*" We use the information from ODP is used to group semantically related pages even if their content is relatively dissimilar. This is done as follows: first, the last level(s) of the hierarchy for categories having more than seven levels are truncated. We assume two web pages to be similar if they share the top seven levels of the hierarchy. Second, clusters containing documents having identical categories are merged. Third, clusters containing pages that have an ancestor–descendant relationship are merged provided that the category associated with the ancestor has at least four levels. For example, clusters with pages having the following two categories are merged: Cluster-1: "*Top/Business/Marketing and Advertising/Internet Marketing/*" and Cluster-2: "*Top/Business/Marketing and Advertising/Internet Marketing/*

Market Research/". In concluding this section, we note that in spite of the advantages of ODP, its utility is typically limited since many pages do not have ODP categories.

4.2.4 Cluster labeling

Each cluster is labeled using the three keywords that have the highest document frequency within it. The query "ant," for example, yields 5 clusters with the following labels: (1) "ants," "insects," "nest;" (2) "directory," "tips," "engine;" (3) "games," "city," "online;" (4) "apache," "related," "tools;" and (5) "user," "software," "management." The use of the labels provides users with a summary of the cluster content and can allow them to rapidly exclude pages that are unrelated to their information need. Additionally, we also use the cluster labels to help users in query reformulation as explained in Section 5.

4.3 Extraction and analysis of location information

The full, unprocessed text from each web page is parsed to extract location information. As a first step of this process, the text is preprocessed to remove all punctuation except for hyphens and forward slashes (since they are used in certain date formats). Next, the country code top-level domain (if available) is extracted from the URL and mapped to its corresponding country. Subsequently, major world region names (such as the "subcontinent", "middle east" or "southeast Asia") are extracted. The document text is then cross-referenced with 3 indices of locations constructed from the same gazetteer (http://www.world-gazetteer.com): (1) countries, (2) states or provinces (for the United States, Canada, United Kingdom, and Australia only), and (3) major world cities. All the extracted location names are saved as a hierarchy of continent, country, state/province, and city. This allows us to define and implement a powerful region-based and point-based model of geographical locations. In this model, cities are defined as point locations and are contained in region-based descriptions such as states/provinces and countries. Thus, interactions with spatial information can be supported using both point-based and region-based queries.

A major factor that degrades precision of location extraction systems is name ambiguity, which consists of two major types: locations with the same name and locations that are also words. For a thorough discussion of location name ambiguity problems, we refer the reader to [2]. We deal with the first issue by looking for a city name only if its corresponding country or state is also found in the document text. We further limit ambiguities by including only those cities which have populations of 10,000 or greater. Major world cities constitute an exception to this rule and are included even if the page does not contain the name of the corresponding country or state. The second issue is dealt with, in a simplified manner, by only treating capitalized terms as potential locations.

4.4 Extraction and analysis of temporal information

The unmodified document texts are parsed to extract their temporal information using regular expressions. Specifically, we look for and extract dates having the following syntax (along with minor variations): (1) "dd/mm/yy | dd/mm/yyyy | mm/dd/yy | mm/dd/yyyy | mm/yyyy | mm/yy | mm/dd", (2) "dd MonthName yyyy | MonthName dd yyyy | MonthName yyyy | MonthName", (3) "yy | numbers preceding "BC"/"AD"/"BCE"/"CE" | numbers following the word "year"", and (4) 4-digit numbers between 1700 and 2099

Formats differing significantly from those indicated above (including those with different punctuation) are not recognized by the time extraction algorithm. In many of

these cases however, the method is still able to extract the year and possibly the month as well. Extracted dates without a year are assigned the year (if there is one) that appears within 13 characters after it. At this point, dates which do not have a corresponding year value are removed, as are redundant dates. As a final step, the validity of dates and years is checked before associating this information with the page.

Since the aforementioned syntax variations comprise the vast majority of date formats, the recall for the date extraction method is quite high. Furthermore, misidentification of a piece of text as a putative date rarely occurs. Consequently, the precision of this method is also very high. The major exception is the last syntax, which sometimes erroneously tags a number in the given range as a year. A quantitative evaluation of these methods is presented in Section 6.3.

5 Supporting experiential user-data interactions

As discussed in Section 1, our design of the proposed interface for supporting exploratory user-data interactions is motivated by the philosophy underlying experiential systems. In it, the information obtained from the various algorithmic components described in Section 4, is displayed in separate but semantically coupled panels as shown in Fig. 2. In this interface the search results are presented as clusters of web pages, and related spatial and temporal information is displayed using an interactive map and a multi-scale timeline. The interface combines the query and presentation spaces and is *direct*. For example, as users mouse-over links in the interface displaying the clustered results, a fourth panel (top right in Fig. 2) presents a thumbnail of the web-page, a summary of its content, and any media files (such as audio or video) that may be associated with it. A user can also choose to click on a cluster. In such a case the thumbnails of all the pages in the cluster are displayed along with any associated location and temporal information in the map and the timeline respectively. These presentation-interaction features allow users to conceptualize the content (either at a page-level or at a cluster-level) using both textual concepts as well as perceptual information. Furthermore, users can also obtain a visual overview of the data in any specific cluster as well as perform *roll-up* and *drill-down* operations. In the following, we describe five classes of interactions that are supported within this interface and their operational semantics. These include: (1) interaction with information present in individual pages, (2) interactions with information present in the clusters, (3) interactions with media-based information, (4) interactions with spatial information, and (5) interaction with temporal information. Note, that the order of using the operators supporting these interactions is in no way constrained and determined solely by the user.

- *Interaction with information present in individual pages:* Two operators are provided to support visualization and interaction with the information present in individual pages. When using these operators, the user interacts with the page-level information present within each cluster. The operators are:

    1. *Page previews*: The goal of page previews is to provide the user with cues that are perceptual and can be assimilated with a low cognitive load. Subsequently, the user may drill-down and obtain details-on-demand. The page preview is initiated through a mouse-over operation. This operator generates the web page's title, preview, summary, URL, and available media files in the page summery panel on the upper right part of the interface. An example is illustrated in Fig. 5.

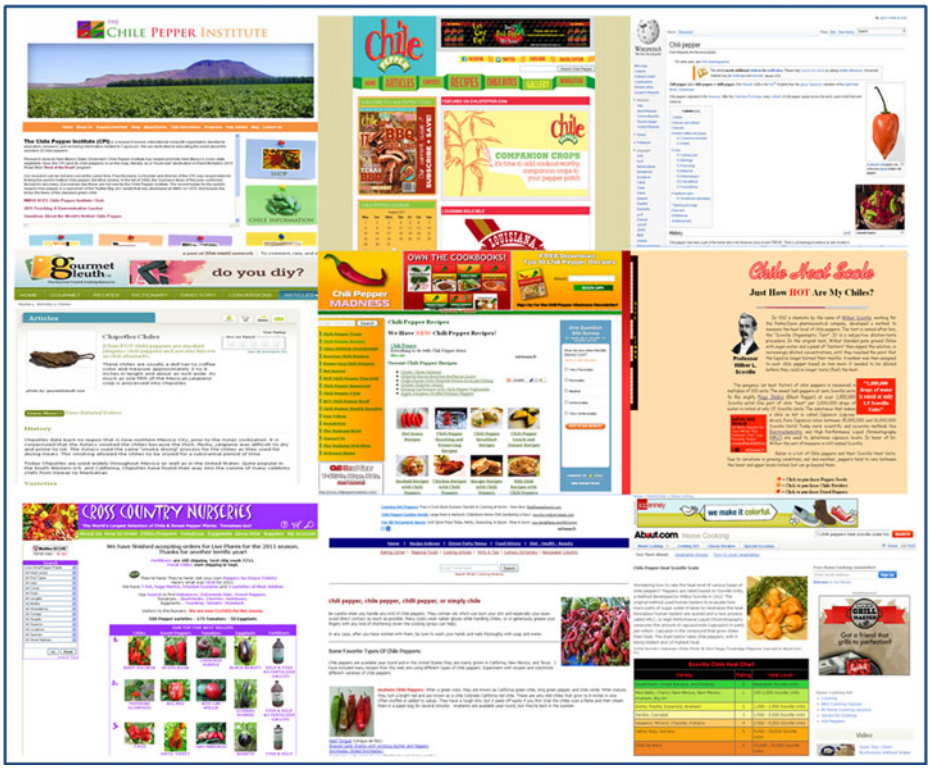


Fig. 5 The cluster preview panel showing page thumbnails of pages from the first cluster corresponding to the query "chili peppers". Users can click on any of the page thumbnails to open the corresponding web page

2. *Opening a page*: The user can open a page either by clicking on the corresponding entry within a cluster or by clicking on the top-right panel where the page previews are presented. This operator follows the standard convention for opening a page as supported in all browsers.

- *Interaction with information present in the clusters*: The system provides operators to obtain overview information as well as details-on-demand both for individual pages as well as clusters of pages. These operators include:

1. *Cluster overview*: Users can select a cluster by left-clicking. The selected cluster is highlighted by a red border to provide a visual cue. Once a cluster is selected, the application provides previews of the pages constituting the cluster (Fig. 5). The preview allows users to rapidly explore a domain without a significant cognitive overload. Further, any spatial and temporal information present in these pages is displayed, respectively, on the map and the timeline.
2. *Cluster information-based query generation*: Right-clicking on a cluster opens a new window where the search engine selected by the user is seeded with a new query that combines the original search term and the labels associated with the selected cluster. The purpose of this operator is to aid the user in query-reformulation by taking into account the information returned in response to the original query as well as the user information need—as indicated by the selection of a specific cluster.

- *Interaction with media-based information*: The media elements within a page are defined to include images, video, audio, as well as PDF documents, MS-Word documents, Power-point documents, and Flash files. When the user selects a page, all the media elements in that page are also displayed through icons under a separate "media" tab. An example is depicted in the screenshot presented in Fig. 2. The user can click the icons to view the corresponding media files.
- *Presentation and interaction with spatial aspects of the information*: The spatial presentation is implemented through the use of OpenMap java toolkit (http://openmap.bbn.com). Cities, whose names occur in the pages, are indicated on the map using circles. Following [47], the size of each circle varies logarithmically with the number of documents containing that location. After the initial display, the user can select areas of interest either through a single click on the location, or by dragging a rectangle across the area. OpenMap converts these mouse events to latitude and longitude, after which the directory of locations is parsed to find extracted locations (cities, states, and countries) that lie fully or partially within the indicated area, or for point selections, within close proximity of the point which was clicked. To disambiguate user intent, the user is then presented with the list of cities, states, and countries thus determined. The user can select one or all of them (see Fig. 8 and Section 5.1). The choice is reflected through highlighting of the documents affiliated with the selected location. Conversely, selecting a cluster of results refreshes the map, displaying only the locations associated with the documents in the cluster. In addition to the aforementioned operators for interaction with spatial information, three cities with the highest frequency of appearance are also listed on top for quicker access. The user can mouse over any one of the top three location hits to see the corresponding state or country information.
- *Presentation and interaction with temporal aspects of the information*: The temporal aspects of the information extracted from the web-pages are displayed using a multi-scale timeline. The initial display shows the distribution of information as a year-level histogram (see Fig. 6). The height of each bar represents the frequency of occurrence of a particular date in the results. The timeline supports multiple time granularities. This allows users to click on a specific year and expand to a month-level view (for that year). The users can recursively click on a month to see the day-level histogram. By allowing users to selectively drill-down on a specific part of the timeline without altering the rest of the timeline, we support the needs for simultaneously maintaining information

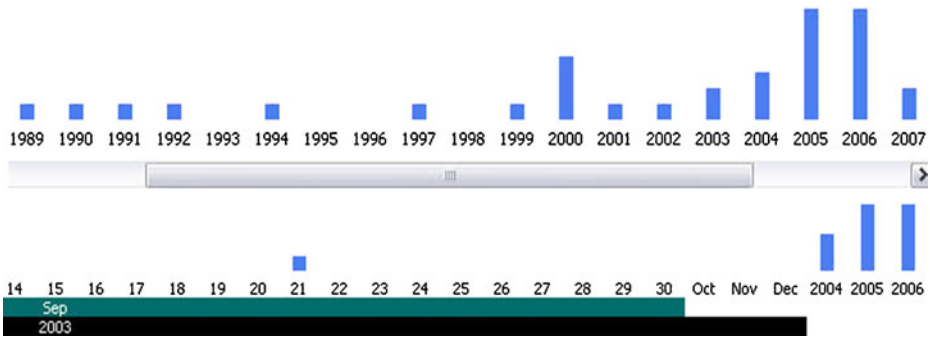


Fig. 6 The multi-scale timeline. *Top*: the default year-level display of information underlying the query term "peanuts". *Bottom*: an instance of multiple temporal granularities supported by the interface. In this case, the user selected to expand the information for the year 2003 to the month level (specifically, the month of September). The timeline shows that that one of the retrieved pages relates to September 21st

overview while allowing detailed focus. Moreover, the interface is kept simple while retaining the ability to present complex multi-scale temporal information. The user can also select a specific date of interest or a region in the timeline corresponding to a time-interval. After the selection, pages containing the selected date(s) are highlighted with a red arrow marking them. Furthermore, these pages are displayed on the preview panel and the locations information in these pages is displayed on the map.

The media, map, and timeline modalities all enable visual overviews of the information both individually and collectively. All these panels support direct manipulations and are reflective, that is, selecting information in any single pane automatically highlights its corresponding characteristics in the others. Users can thus interact, experience, and explore the information using any of the available perspectives (thumbnail previews, media, text-based semantic similarity, spatial characteristics, and/or temporal characteristics).

5.1 Experiential user-data interactions: a case study

In this section, we present a case study that demonstrates an integrative use of the various modalities as well as the experiential interactions supported by the system using screenshots. The query in this case consisted of the term "earthquake". The initial display of the data in terms of the different semantic perspectives is shown in Fig. 7 with the various clusters shown in the cluster panel. The screenshot shows the situation where the user had selected to preview the constituent pages in the first cluster. The thumbnails of the pages from this cluster can be seen in the preview panel on the top right. The user next decided to interact with the information using the spatial display. Specifically, a region of interest along the US west coast was selected as displayed in the screenshot shown in Fig. 8 (top). In Fig. 8 (bottom) the system response is captured: a location box was displayed so that the user could indicate a specific location of interest within the selected region. The user responded by selecting the city of San Francisco as the location of interest. The system updated the information by displaying the pages from the first cluster that were related to San Francisco. This can be partly seen in the screenshot in Fig. 9 (top). The reader may note the pages in the preview have changed to only include those that have the location "San Francisco". Next the user switched to interacting with the temporal aspects of the data by scrolling the timeline to the early years of the 20th century. The histogram in Fig. 9 (top) shows a number of hits containing the year 1906. The user next clicks on this year to get a month level view of the data. This is illustrated in the bottom screenshot in Fig. 9. The resultant information is reflectively updated in the other panels; the reader may note among others the links highlighted with arrows on the right and the pages in the preview. Finally, the user switched to visual selection from the preview and selected a page on the 1906 San Francisco earthquake by clicking on the corresponding thumbnail.

6 Experimental evaluation of the system

In this section we present results from experiments that evaluated the components of the proposed system. The evaluations targeted the following aspects: (1) effectiveness of grouping semantically related content, (2) quality of labels assigned to the clusters, and (3) the accuracy of spatial and temporal information extraction, when compared with manually extracted ground truth.

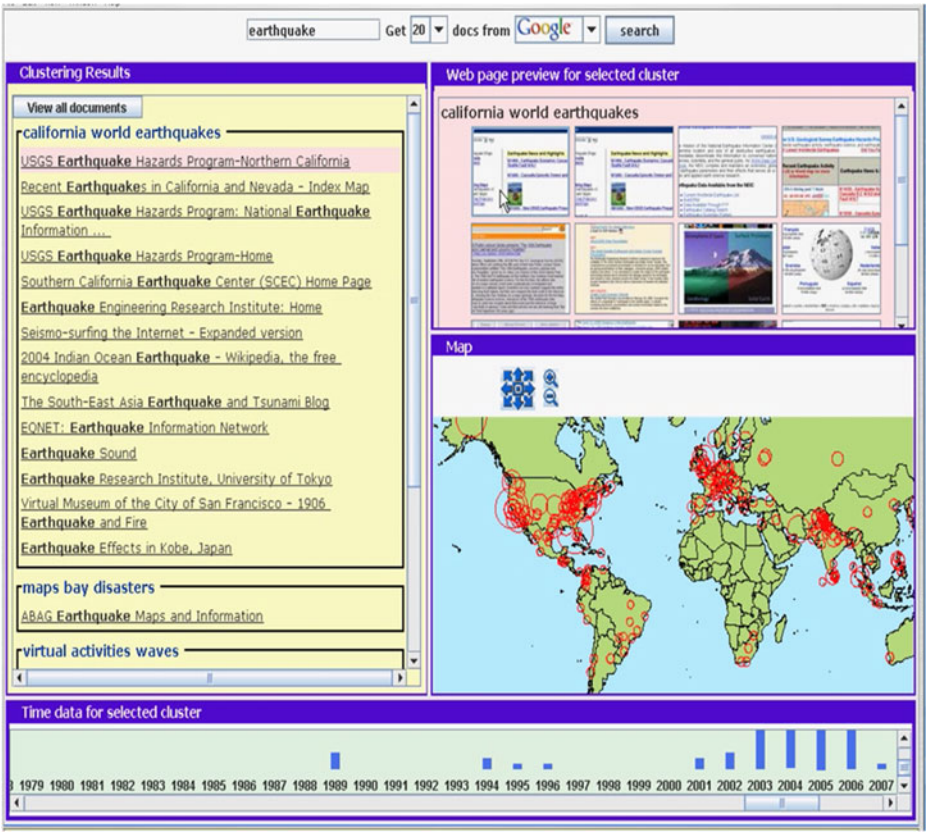


Fig. 7 Case study involving the different interaction modalities and their integrative usage for the query "earthquake". The results panel previews the pages constituting the first clusters. The map shows a geographical distribution of the hits and their temporal distribution is shown on the timeline. Previews of specific pages in the cluster can be obtained through a mouse-over operation. Figures 8 and 9 show the subsequent steps in the case study

6.1 Effectiveness of grouping semantically related content

We note that given a query, the key goals of clustering web-pages include reduction in the cognitive load by decreasing the number of links a user needs to peruse and providing the user an understanding of the informational variability. Therefore, we evaluated the system vis-à-vis these objectives. To assess performance with respect to the first objective, we counted the number of clusters as an indicator of the reduction in data size. Success in attaining the second objective was determined by appraising the correctness of the clusters using the mutual information measure. This measure was used to quantify the agreement between algorithmically-derived clusters and manually defined ground-truth clusters.

In the first experiment we considered the 40 queries described in Section 4.2.1 and analyzed the top 20 documents retrieved for each query. A dynamic background set of 50 documents was used with an empirically-determined first-stage clustering threshold of 0.90. Results were generated for LSA + TFIDF both with and without category-based clustering. This experiment was designed to quantify the advantages (if any) due to the ODP-driven clustering. In it, the amount of data reduction was calculated as the percent difference

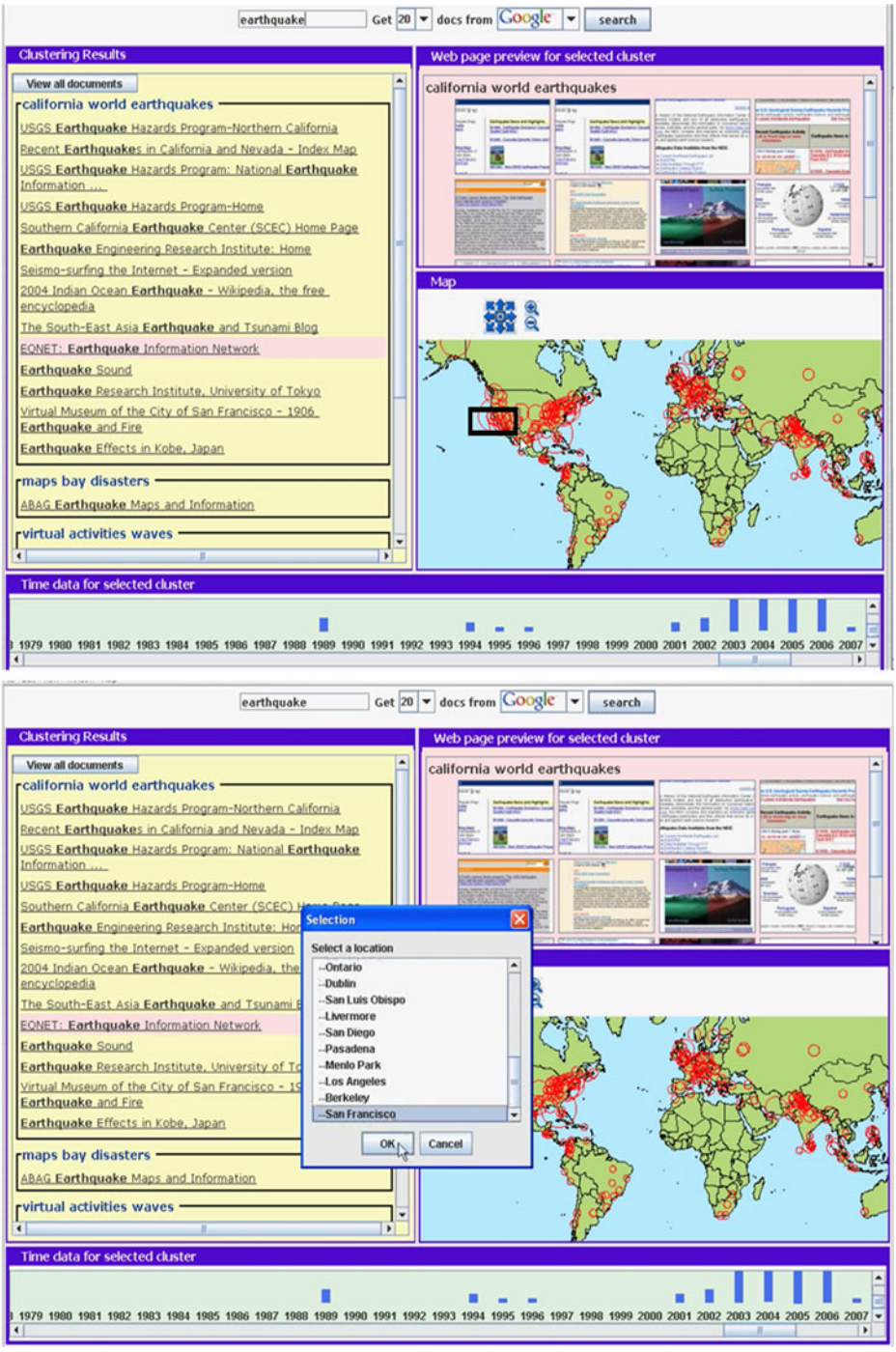


Fig. 8 The user chooses to interact with the spatial aspects of information. The interactions are initiated (*top*) by highlighting a geographical area of interest around the west coast of the US. (*Bottom*) A location box appears to help in disambiguation and the user selects San Francisco as the location of interest

between the number of clusters and the number of documents they contained. The results for this experiment are shown in Table 5. These results indicate that LSA + TFIDF-based clustering led to an average of 54% data reduction while clustering using LSA + TFIDF along with category information from ODF led to 62% data reduction. Clearly, such a decrease in data volume is advantageous in reducing the cognitive load of the user provided that the clustering of the documents is accurate (that is, if the clusters contain semantically related pages).

The second experiment was designed to investigate the quality of clustering by assessing the extent to which the clusters contained related information. This evaluation was conducted for clusters obtained under three conditions: by using LSA, by using LSA + TFIDF, and by using LSA + TFIDF along with ODP category information. The cluster quality for each of these cases was quantified using the [0,1]-normalized mutual information measure [45]. This measure is defined in Eq. (7) and unlike other cluster quality measures such as purity or entropy, does not favor small clusters.

$$\varphi^{(NMI)}(\lambda,\kappa)=\frac{2}{n}\sum_{l=1}^{k}\sum_{h=1}^{g}n_l^{(h)}\log_{k\cdot g}\left(\frac{n_l^{(h)}n}{n^{(h)}n_l}\right) \tag{7}$$

In Eq. (7), λ denotes the algorithmically obtain clustering, κ the true clustering (ground truth), k the number of algorithmic clusters, and g the number of true clusters. The term n_l denotes the number of objects in cluster l according to λ, $n^{(h)}$ is the number of objects in cluster h according to κ, $n_l^{(h)}$ is the number of objects in cluster l according to λ and in cluster h according to κ, and n is the total number of objects. The formulation of mutual information described in Eq. (7) is symmetric in terms of λ and κ. A mutual information value of 1 indicates perfect clustering, while random clustering gives a value of zero in the limit. It should be noted however, that the best possible labeling leads to a value of less than 1, unless the classes are balanced. Unfortunately, alternate formulations that can lead to the value of 1 for best possible labeling tend to be biased [45]. It should also be noted that data reduction and mutual information are mutually exclusive criteria. This is due to the fact that the greatest cluster coherence is achieved when clusters are small (most notably in the trivial case, of singleton clusters). However, in such a case there is very little data reduction. Conversely, having a few large clusters would lead to significant data reduction, but the resultant clusters would generally contain diverse information and therefore have low mutual information values. Thus, data reduction and mutual information values must not be interpreted in isolation.

As is apparent from Eq. (7), use of the mutual information measure requires the definition of ground truth. A clear definition may however, not be possible in every situation since complex information may be interpreted (and clustered) differently by different people. Consequently, for this experiment we used the committee-voting approach to define the ground truth as outlined in Section 4.2.1. The committee was comprised of the three authors. For each of the queries, the top 20 hits were independently analyzed and clustered by each committee member. The final ground truth was defined by majority voting. The idea behind this approach was to favor clusters that were agreed upon by a majority.

In Fig. 10, we graphically present the mutual information values for clusters obtained with LSA, LSA + TFIDF, and LSA + TFIDF along with ODP category information. The results presented in Fig. 10 show that the addition of TFIDF did indeed improve the mutual information of the final clusters in most cases, on average by about 0.03 units or roughly about 3.2% given the range of observed mutual information values. However, for six

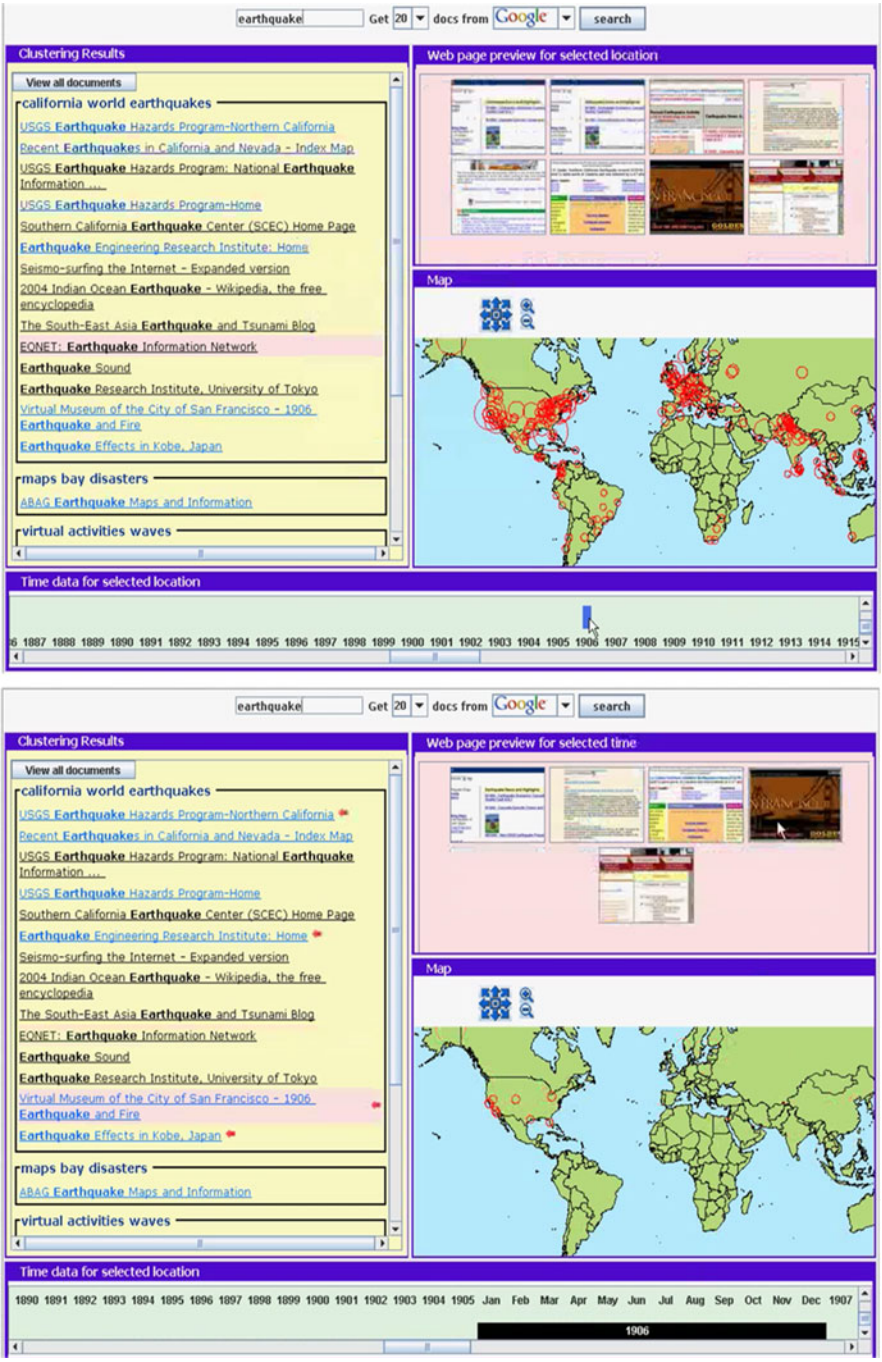


Fig. 9 (*Top*) The user switches to interacting with the temporal aspects of the information by scrolling the time-line to the early years of the 20th century. A significant number of hits corresponding to the year 1906 can be observed in the top snapshot. The user next clicks on this year to expand to the month view (*bottom* snapshot). The resultant hits are updated in the other panels. The user then identifies a specific page of interest corresponding to the 1906 San Francisco earthquake

queries (query numbers: 4, 6, 16, 21, 22, and 36) the mutual information decreased for clusters obtained using LSA + TFIDF as compared to clusters obtained with LSA alone. For all these queries, the results included pages which either had insufficient text or contained information (such as addresses) which was not useful for clustering. In [45] for sample size $n=800$, mutual information values around 0.4 to 0.5 were considered to be excellent. While the sample size was much smaller in our case, it is still interesting to note that for all queries the mutual information for the clustering (using LSA + TFIDF as well as LSA + TFIDF along with ODP categories) exceeded (often significantly) the value of 0.5. Furthermore, for clustering involving ODP information, the mutual information values were, with rare exceptions, significantly higher. This implies that the clusters found algorithmically often agreed with manual interpretation of the information. The exceptions, when ODP categories were used, were decreases of about 0.1 in the mutual information for the 9th and 10th queries (corresponding to "jobs in bay area" and "stock car racing") and minor decreases for the 19th and 25th queries ("ODP" and "computer", respectively). One of two reasons typically contributed to such results: first, in certain cases (such as for the queries "stock car racing", "ODP", and "computer"), the original clusters had diverse content. When merged, an even more heterogeneous cluster was created. Second, sometimes original clusters that were homogeneous were incorrectly merged (this happened for the query "jobs in bay area"). In certain other cases, the use of ODP categories did not lead to any change in the mutual information. This was either because the pages in the same category were already clustered together or because the pages did not have ODP classifications.

Finally, it is instructive to analyze specific queries, the retrieved results, and the definition of the corresponding ground truth clusters to understand the complexity of interpreting and assessing clustering of web-search results. As an example, for the 17th query (peanuts), two of the authors placed "peanut products" and "recipes" in separate clusters, while one of them placed these pages together. Based on the committee voting, these pages were consequently placed in separate ground truth clusters. The system however, placed these pages in the same cluster leading to a relatively low mutual information value for this case. The score would have been higher if the pages were in separate ground truth clusters.

6.2 Quality of cluster labeling

The cluster labels play an important role in the proposed system both by providing cues to the underlying content as well as in query-reformulation. Thus, the quality of these labels is important. In this section, we present results from a case study involving five randomly selected queries. These results are meant to provide the reader an intuition about the nature of the labels, as determined by our method, and the extent to which these labels represent the information of the corresponding cluster (s). The queries and the corresponding cluster labels are shown in Table 6. The table also contains manually generated interpretation of the cluster contents. The results show that labels tend to be descriptive and accurate in the vast majority of cases. Since clustering and labeling of clusters is particularly beneficial for polysemous queries, we briefly analyze the cluster labels for the query "peanuts". The two largest clusters, in this case, contained 7 pages each. The first of these clusters contained information about the comics-series Peanuts and was labeled with the keywords "charles", "snoopy", and "complete". The second cluster contained results about the legume peanut and was labeled using the terms "recipes," "nuts," and "products." Most of the remaining clusters

Table 5 Evaluation of the reduction in the information presented to the users through clustering. Two scenarios are evaluated. In the first case, the clustering is purely data-driven and based on LSA + TFIDF. In the second case, the clusters produced in the previous step are further refined using category information (abbreviated as CAT in the last column)

Query no.	Query	LSA + TFIDF		LSA + TFIDF with CAT	
		No. of clusters	% data reduction	No. of clusters	% data reduction
1	eclipse	8	60	7	65
2	phone	8	60	6	70
3	image	10	50	7	65
4	New York	11	45	6	70
5	wolf	8	60	6	70
6	ant	9	55	5	75
7	brother	9	55	9	55
8	sports car	9	55	6	70
9	jobs in bay area	15	25	13	35
10	stock car racing	11	45	7	65
11	nano	7	65	5	75
12	vista	8	60	7	65
13	Amazon rainforest	14	30	12	40
14	Japanese comics	13	35	13	35
15	distance teaching	9	55	9	55
16	tornado	12	40	10	50
17	peanuts	8	60	7	65
18	flowers	8	60	8	60
19	ODP	9	55	6	70
20	security	8	60	6	70
21	Dali	6	70	3	85
22	fall festival	12	40	12	40
23	notebook	11	45	10	50
24	Hilton in Paris	11	45	10	50
25	computer	11	45	10	50
26	Sony	7	65	6	70
27	jaguar	6	70	5	75
28	time	7	65	5	75
29	eraser	6	70	5	75
30	pain relieve	15	25	14	30
31	Mozart	7	65	6	70
32	paper clip	8	60	8	60
33	DDR	4	80	4	80
34	rent	9	55	8	60
35	chili pepper	8	60	4	80
36	diamond	12	40	12	40
37	woodstock	9	55	8	60
38	treasure island	9	55	7	65
39	panda us	10	50	10	50
40	chocolate factory	7	65	5	75
Average		10	54	8	62

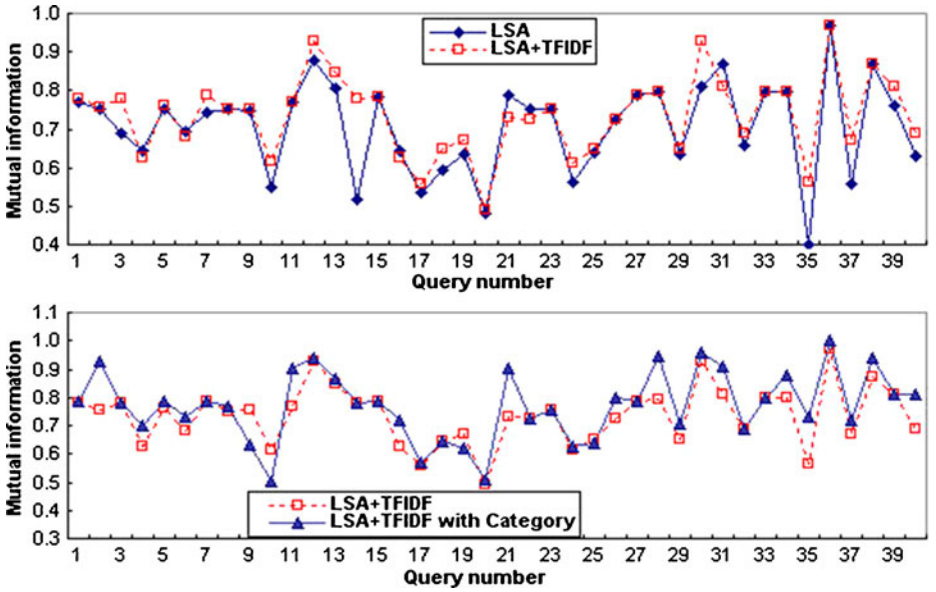


Fig. 10 Mutual information values corresponding to different clustering methods

were singletons, and were labeled with keywords extracted from the document titles. For the other four queries in Table 6, the labeling was mostly consistent with the cluster content. The exceptions were the second and third clusters of the query "eclipse". Both these clusters contained a single page each. We followed-up these two cases and found that the titles of these documents failed to precisely describe their content, leading the algorithm to assign inconsistent labels.

6.3 Effectiveness of spatial and temporal information extraction

The effectiveness of the location and time extraction modules was evaluated by comparison with manual extraction on the set of 40 queries described in Section 4.2.1. For each of the queries, the first 20 results were analyzed. As part of the analysis, all explicit location names and dates in the pages were manually tagged and compared with the automatically extracted information. The following were considered to be unambiguous spatial and temporal informational attributes: (1) country names, (2) state/province names (for the aforementioned countries), (3) major world cities, (4) cities accompanied by the name of the corresponding country or state/province, and (5) years (with or without month/day information). The results of the comparison are presented in Table 7. On this data set, the recall value for location name extraction was approximately 93% while the recall value for date extraction was approximately 96%. This indicates that the vast majority of location and time-related information was identified. The precision for date and location name extraction was close to 99% and 88% respectively. The lower precision for location extraction (as compared to date extraction) was due primarily to name ambiguity. By contrast, the high precision and recall for date extraction corresponded to the fact that the algorithm was designed to detect the vast majority of date formats and was rarely incorrect. The results

Table 6 Labels generated for clusters corresponding to various queries along with manual interpretation of the cluster contents

Query	No.	Cluster labels	No. of docs	Description of documents in cluster
Eclipse	1	development software java	8	software development & download
	2	items available planning	1	Music order catalog
	3	create easy students	1	Crossword puzzle
	4	students years clues	1	Students developed game
	5	sun eclipses solar	5	Information about solar/lunar eclipse
	6	general software	2	Software game
	7	business future news	2	Business solution
Ant	1	ants insects nest	11	Information about insect ant
	2	directory tips engine	1	Directory search engine
	3	online games city	1	Online games
	4	apache tools related	4	Project/discussion related to apache
	5	user software management	3	Software development/management
Phone	1	pages directory telephone	6	Yellow page at directory search
	2	business service internet	10	Telephony service, voip
	3	reverse people codes	1	Reserve phone directory
	4	number services toll	1	Phone number spell service
	5	yellowpages pages business	1	Yellowpage for business & people
	6	service communication calling	1	Wireless communication service
Peanuts	1	schulz charles comic	7	Information about the comics peanuts
	2	software homepage	1	Peanut software download
	3	lee bros boiled	1	Shopping for Lee Bros boiled peanuts
	4	character quizzilla	1	Mental quiz
	5	crakerjacks guide	1	Peanuts and crakerjacks bank
	6	recipes butter nut	7	Peanuts products and recipes
	7	coup files yahoo	2	Music video and flash file by group X
Security	1	advisory latest computer	7	Computer latest virus report/advisory
	2	national agency central	1	National Security Agency
	3	antivirus intrusion Chinese	2	Information about antivirus software
	4	administration documentnews	2	Social security online and news
	5	browser main	1	Netscape browser download
	6	internet systems worldwide	1	Internet security system
	7	homeland contact threat	6	Information about homeland security

obtained by the proposed method are comparable with those from [58], where a multi-stage approach was proposed for locating geographic information from web pages. In [58], test sets of 500 pages each were constructed based on region and language. For these sets precision in the range of 86% to 93% and recall in the range of 94% to 98.5% were reported.

Table 7 Precision and recall values for identification of location and time related information in the web pages

	Date extraction	Location-name extraction
Precision	99%	88%
Recall	93%	96%

7 User study and evaluations

7.1 Study design

The purpose of the user study was to evaluate how well the proposed paradigm helped participants on searching and exploring information compared to some of the most commonly used commercial systems. In particular, we focused on how participants utilized each of the modalities in our system and whether the efficacy of the proposed paradigm improved as users got better acquainted with it over time. The study involved both quantitative and qualitative evaluations. It consisted of two sections and was spread over a week; participants were asked to complete the first section at the beginning of the week and the second section at the beginning of the following week. In between the two sections, participants were asked to continue using our system in order to further familiarize themselves and gain expertise with its different features. The purpose of the extended study was to estimate any improvements in how effectively the participants employed the system over a prolonged period of use.

Two sets of questions were interchangeably used to eliminate any bias associated with ordering. Each questionnaire included 20 tasks where each task required participants to issue a query and interact with the retrieved information using the designated system (s). The time and number of clicks needed to find the desired information was recorded for each query executed on each of the four systems for every participant. Of the total set of queries, 14 queries were compared using our system, Google, Vivisimo, and Grokker while 6 queries compared our system with Vivisimo. The information goals behind these queries were highly non-trivial. In Table 8, we list some of the queries and the specific aspects of information they were designed to be most related to. For the above information goals, we anticipated 36% of the queries to be text oriented, 21% to be spatial in nature, 18% to be temporal, and 25% to be media oriented. It should be noted that information goals could also be satisfied by exploring aspects of information different from those anticipated in the design of the experiment. For instance, the media oriented information goal of finding a PDF brochure on turtle-bay, Hawaii, could alternatively be satisfied using spatial cues and search.

Twenty participants were recruited for the study. The age of these participants varied between 21 and 45 years. The group consisted of 11 female and nine male participants.

Table 8 Example information goals used in the first part of the user study

Information aspect	Example information goals
Media oriented	Find: (1) video on recent immigration protests in the US, (2) video of the movie *Heaven's Gate*, (3) PDF tutorial on SPSS, (4) PDF brochure on *turtle-bay (Hawaii),* (5) piano music score *Fur Elise* by Beethoven
Temporal	Find the date of (1) IKEA's opening in Taipei (2) release of LINUX, (3) first discovery of dinosaur eggs, (4) opening of the channel tunnel between UK and France (5) the year in which the hot-air balloon was invented
Spatial	Find the location of: (1) places where Pandas can be found in the US, (2) Asian country where mummies were discovered (3) closest golf course from San Francisco with an ocean view (4) City in which Microsoft Inc opened its first Asian division (5) PDF brochure on *turtle-bay (Hawaii)*
Text oriented	Find: (1) Five hottest varieties of chilli-pepper, (2) Five uses of lavender in cooking and medicine (3) what Bruce Lee studied in Univ. of Washington, Seattle (4) the religion of which the red lotus is a symbol, (5) recipe for fortune cookies

Four of the participants were undergraduate students and five were graduate students. The remaining 11 participants had varied educational backgrounds and were drawn from different professions. These participants were divided into 4 groups. Each group ran queries in different order on four applications based on the Latin Square to avoid bias owing to the order in which the tasks were performed. This was important, since participants could have had a sense of the information content pertinent to a query after they had run it on the first system. Therefore, without such precautions, the results would have favored the last three applications being evaluated. The data from the user studies was used to test the following hypotheses:

1. Participants would experience greater success and satisfaction with the proposed paradigm (as embodied by our prototype system) than with other commercial search engines.
2. Compared with Vivisimo, which supported clustering of results in a manner somewhat analogous to our approach, the proposed system would give participants a better insight into the informational structure underlying a query in a shorter amount of time.
3. Over time, as participants become comfortable with the proposed paradigm, the effort and time need to obtain the desired information, will decrease.

We used pair-wise comparison of different variables to analyze the affect of different factors involved the study. Analysis of Variance (ANOVA) was used to determine the statistical significance of the difference among the mean scores of two or more variables. When the p-value was less than 0.05, the two variables were considered to be significantly different.

7.2 Comparison of the systems in terms of time and number of clicks

For each query, we compared the time and number of clicks needed to find the desired information across the proposed system and three alternatives: Google, Vivisimo, and Grokker. As shown in Fig. 11, across both these metrics, our system was more efficacious than the other three systems. That is, on an average, participants were able to reach their information goals at least 20 s faster or 6 fewer clicks than all the three commercial systems. The effect of participants on the time needed to satisfy the information goal was significant for each system tested: *Google* (F=6.43, p<<0.001), *Vivisimo* (F=4.35, p<< 0.001), *Grokker* (F=2.425, p<<0.001). Even though our system showed variability between participants (F=1.82, p=0.018), all participants were able to reach their information goal in less time using our application, which displayed considerably lower values for the standard deviation compared to the others (the p-value for our system was greater than others). On the other hand, the affect of participants was not significant in terms of number of clicks (p<<0.004) across all the applications investigated. The reader

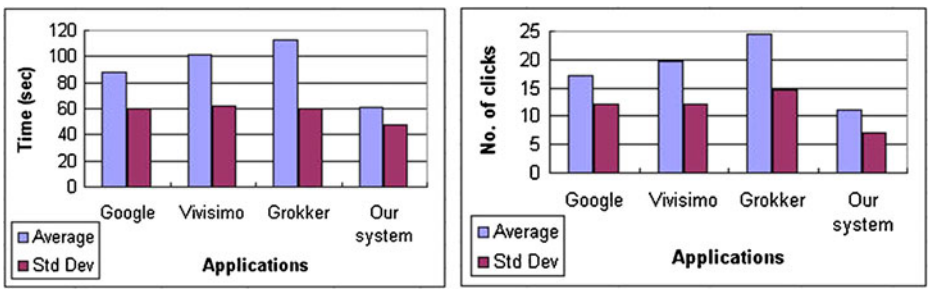


Fig. 11 Average time and number of clicks required to satisfy the information goals on each system

may also note the high variance observed in this experiment. Two slowest users were found to take almost twice the amount of the time and number of clicks than the fastest user, thereby affecting the mean and standard deviation values considerably. It may be noted that our observations are similar to those reported in [6], where a similar variability in search times was observed across users.

7.3 Assessment of exploratory capabilities

A key challenge in web-search is that of "information discovery". While such an idea is very intuitive, its formal definition is complex and arguably subjective. However, it is reasonable to associate this notion with the following two characteristics: (1) obtaining any information that is broadly related to the query, whose relation to the information goal was not specifically used as part of the query. (2) Understanding the semantically relevant and related groups of information relevant to the query. Intuitively, these characteristics correspond to how the information is structured with respect to the query. For instance, the query "treasure island" can lead to information that corresponds to "Treasure Island the movie", "Treasure Island, the city", "Treasure Island the book", and "Treasure Island the resort/casino in Las Vegas".

To evaluate our research in this context, we studied it using six information goals. These goals were related to the following topics: (1) exploration of the professional history of Dr. Chung-Sheng Li (a scientist at IBM research), (2) researching the animation movie Snow White, (3) discovering events in the life of the Mexican painter Frida Kahlo, (4) researching the users favorite TV show, (5) exploring the life of Arnold Schwarzenegger, and (6) researching volcanic eruptions. As part of this experiment, participants were asked to explore information relevant to these six information goals using the proposed system and Vivisimo. To limit the time requirements, the participants were also asked to spend no more than 3 min on each task. At the end, the participants rated the systems on a scale of 1 to 5, where a score of 1 corresponded to the notion "very difficult" and a score of 5 to the notion "very easy". The capabilities of the system both for information discovery as well as for understanding the structure of the underlying information were evaluated. The results for this study are presented in Fig. 12 (left). The proposed system scored an average of 4.33 on ease of both information-discovery and understanding of information structure while Vivisimo got scores of 3.2 and 3.3 respectively. The participants could only find the answers to 49 out of 240 tasks (six queries × two sections × 20 users) in 3 min by using Vivisimo. This implied a completion rate of 20.4% for tasks using Vivisimo. On the other hand, users were able to find the answers to 102 tasks (42.5% of the total tasks) within 3 min by using proposed system.

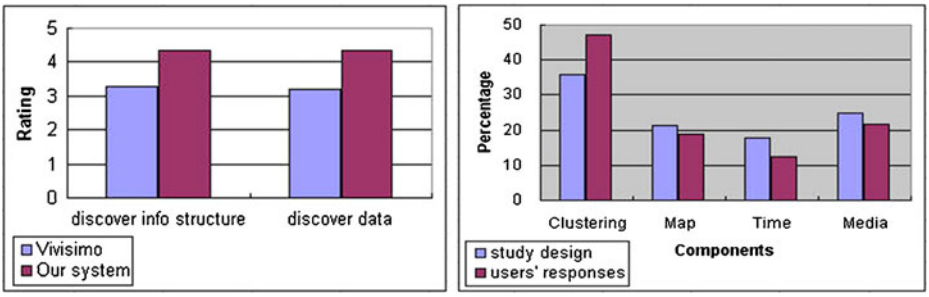


Fig. 12 *Left*: user ratings for Vivisimo and the proposed system for information/data discovery tasks. *Right*: expected and actual usage of the specific interaction modalities during information search and exploration. In this figure, "clustering" denotes page content-based clustering, "map" denotes spatial interactions, "time" denotes temporal interactions, and "media" denotes media-based interactions

7.4 Usage of different user-data interaction modalities

Given the various perspectives on the information supported in our approach, an interesting question is how important users found each module as they searched for specific information. To evaluate this, user responses on all the 28 information goals used as part of the user study (14 queries per session) were analyzed. As mentioned earlier, from our perspective as study designers, the necessary information relevant to 36% of the queries could best be retrieved using the text-driven clustering, 21% could be addressed using spatial information, 18% using temporal information, and 25% using the media display.

At the conclusion of each query, the 20 participants were asked to name the specific interaction modality (text-clustering, time, location, or media-display) that was most helpful to them in finding the relevant information. These results are shown in Fig. 12 (right). The responses were text-clustering (47%), spatial-display (19%), temporal-display (12%), and media-display (22%). The choice of text-driven clustering for a greater number of problems than what was anticipated by us is noteworthy. Discussions with the participants helped identify two reasons for this skew. First, due to the historical predominance of text-based web-information retrieval, some users were more inclined to use textual modality to find the necessary information. Second, some users preferred the clustering because it allowed them to understand the overall information structure and yet rapidly drill down to the necessary information. It is also important to note, that information goals that involved finding and retrieving media, took the least amount of time to be fulfilled across the majority of users. Typically, for these goals, the participants were observed to narrow down the required information using the various interaction modalities and use the media tab in the final step to access the data.

7.5 User adaptation over time and subjective rating of systems

In this section, we analyzed if and how the brief but longitudinal nature of the study was reflected in the usage statistics. In the first section of the study, participants spent, on an average, 89 s on Google, 105 s on Vivisimo, 114 s on Grokker and 66 s on our system to reach the desired information goal (we remind the reader, the complex nature of the involved queries). With respect to these times, in the second stage, participants took less time across all the applications. The average percentage reductions in times were: 2.7% for Google, 5.9% for Vivisimo, 1.9% for Grokker and 13.3% for the proposed system (Fig. 13). The fact that the proposed approach required less time-to-information amongst all the

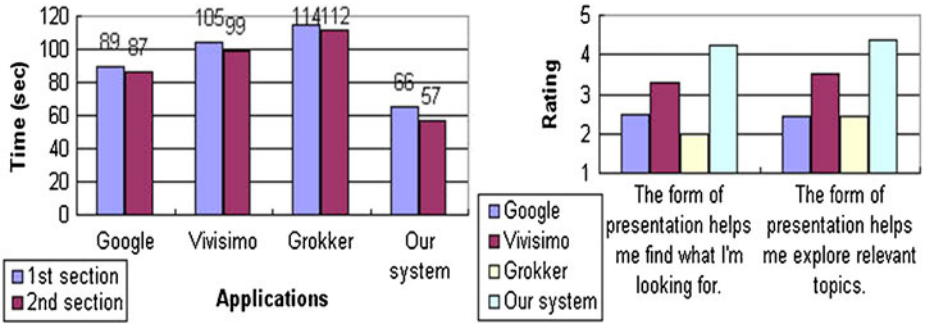


Fig. 13 Time to reach the information goal in the first and second sections of the study which were separated by a week (*left*). Ratings assigned to the four systems by users (*right*)

systems in both stages, illustrates the applicability of alternate paradigms to mediate user-data interactions in web search, especially for complex exploratory queries.

In the second part of this experiment, the users rated each application in terms of how well the corresponding presentation of results facilitated search and assimilation of information. The rating was on a Likert scale of 1 (strongly disagree) to 5 (strongly agree). Our application received the highest scores of 4.1 and 4.25 respectively, on the ratings of the two criteria: helpfulness of finding desired information and exploring relevant data. These results are graphically shown in Fig. 13 (right). In this experiment Google was rated around an average of 2.5 on both these criteria. The users rated Vivisimo at 3.3 in terms of helpfulness in finding the desired information and 3.5 for supporting data exploration. These results indicate that some form of semantically relevant clustering of search results is more effective in helping users find information, than simple page ranking, especially for complex and exploratory information needs.

While the results of the user study point to the efficacy of the proposed approach, few limitations should be noted. First, the current system uses a fixed set of perspectives. These perspectives may not be optimal in certain situations. For instance, certain web pages may not contain sufficient temporal or spatial information to allow users to take full advantage of the timelines and the map. A mechanism that dynamically invokes other, more suitable perspectives, for such cases may be desirable. Second, in this study a small sample of users from a single location was involved. While the gender ratio of the users was nearly balanced, the age of the users ranged between 21 years and 45 years and may not be fully representative of the age of all web users. Finally, cultural differences influence how well the results of a study can be generalized to other populations [23]. These observations should be kept in mind while interpreting or extrapolating our results to another population.

8 Discussions and conclusions

This paper presents a detailed description of our research on designing an experiential user-data interaction environment for exploratory web search. The underlying design paradigm seeks to support human-machine synergy by identifying semantic correlations in the retrieved information and facilitating direct interactions between the users and the data. Further, mechanisms such as spatial and temporal displays, semantic clustering, and tight coupling between different views of the data help maintain user state and context and provide insights into the relationships within the information.

The approach proposed in this paper differs from prior research both in terms of design philosophy (as described in Section 3.1) and technical details. From the perspective of clustering web search results, a key difference of the proposed work from other methods is that it combines page content-based algorithmic clustering with manually derived ODP clustering. Further, it uses a novel term weighting strategy by combining latent semantic analysis with TFIDF weighting. In the proposed approach, we also analyze the page content for perceptually important cues, such as media files, geographical information, and temporal information and use them to provide alternate perspectives on the underlying relationships within the data.

From a visualization and interaction perspective, the proposed approach does not employ any abstract presentation or interaction metaphors, such as glyphs [5, 33] or sketches [26]. Methods such as [12, 13, 19], support exploration by visualizing the term distribution in the retrieved pages relative to the terms in the query. By contrast, our visualization of the information is not restricted by the query terms. Consequently, the user has greater opportunities for exploration, serendipitous discovery, and assimilation

through the use of different semantic perspectives. The supported interaction operators are also simple, involving only point-click, select, and drag. Yet, these operators allow powerful interactions with the semantic, spatial, temporal, and visual aspects of the information. Finally, the proposed approach represents a novel integrative visualization, where in real-time, a user is provided with a semantically correlated combination of clustering, textual and visual cluster overviews, page previews, media previews, interactive geographical information presentation, and interactive multi-scale temporal information presentation.

Results from investigations involving detailed user-studies and evaluations conducted in comparative settings with other solutions underline the efficacy and value of the proposed paradigm both in information retrieval and information exploration tasks. While the results in this paper were obtained in the context of web search, they are expected to be of relevance to a wide class of problems in information retrieval involving multifarious heterogeneous data and complex information needs.

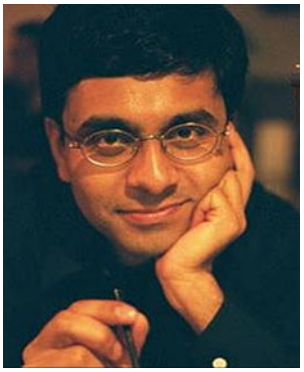
Acknowledgments The authors also thank Dil Chiture, Liu Yang, and Wen-Cheng Sun for participation in early parts of this research, results from which were published in [39]. The authors also thank the anonymous reviewers for their comments which led to many improvements in the paper. Finally, RS would like to thank Ramesh Jain for introducing him to the ideas underlying experiential computing.

References

1. Ahlberg C, Shneiderman B (1994) Visual information seeking: tight coupling of dynamic query filters with starfield displays. In: Proceedings of the SIGCHI conference on Human factors in computing systems: celebrating interdependence (CHI'94). ACM, New York, NY, USA, pp. 313–317
2. Amitay E, Har'El N, Sivan R, Soffer A (2004) Web-a-where: Geotagging Web Content. In: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'04). ACM, New York, NY, USA, pp. 273–280
3. Cai D, He X, Li Z, Ma W-Y, Wen J-R (2004) Hierarchical Clustering of WWW image search results using visual, textual, and link information. In: Proceedings of the 12th annual ACM international conference on Multimedia (MULTIMEDIA'04). ACM, New York, NY, USA, pp. 952–959
4. Carpineto C, Osinski S, Romano G, Weiss D (2009) A survey of Web clustering engines. ACM Computing Surveys Vol. 41, No. 3, Article 17
5. Chau M (2011) Visualizing web search results using glyphs: design and evaluation of a flower metaphor. ACM Transactions on Management Information Systems Vol. 2, No. 1, Article 2
6. Dumais S, Cutrell E, Chen H (2001) Optimizing search by showing results in context. In: Proceedings of the SIGCHI conference on Human factors in computing systems (CHI'01). ACM, New York, NY, USA, pp. 277–284
7. Ferragina P, Gulli A (2005) "A personalized search engine based on web-snippet hierarchical clustering." In: Special interest tracks and posters of the 14th international conference on World Wide Web (WWW'05). ACM, New York, NY, USA, pp. 801–810
8. Gemmell J, Bell G, Lueder R, Drucker S, Wong C (2002) MyLifeBits: fulfilling the Memex vision. In: Proceedings of the tenth ACM international conference on Multimedia (MULTIMEDIA'02). ACM, New York, NY, USA, 235–238
9. Grosky W, Sreenath DV, Fotouhi F, Wietrzyk V (2006) The multimedia semantic Web. Int J Comput Sci Eng 2(5/6):326–340
10. Havre S, Hetzler B, Nowell L (2000) ThemeRiver: visualizing thematic changes in large document collections. IEEE Trans Vis Comput Graph 8(1):9–20
11. Hearst MA (1995) TileBars: visualization of term distribution information in full text information access. In: Proceedings of the SIGCHI conference on human factors in computing systems (CHI'95). ACM Press/Addison-Wesley Publishing Co., New York, NY, USA, pp. 59–66

12. Hoeber O, Yang XD (2006) The visual exploration of web search results using HotMap. In: Proceedings of the conference on Information Visualization (IV'06). IEEE Computer Society, Washington, DC, USA, pp. 157–165
13. Hoeber O, Yang XD (2008) Evaluating WordBars in exploratory web search scenarios. Inf Process Manag 44(2):485–510
14. Hsu YW, Moon N, Singh R (2006) Designing interaction paradigms for web-information search and retrieval. In: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI'06). IEEE Computer Society, Washington, DC, USA, pp. 815–822
15. Jain R (2003) Experiential computing. Commun ACM 46(7):48–55
16. Kammerer Y, Gerjets P (2010) How the interface design influences users' spontaneous trustworthiness evaluations of web search results: comparing a list and a grid interface. In: Proceedings of the 2010 Symposium on Eye-Tracking Research and Applications (ETRA'10). ACM, New York, NY, USA, pp. 299–306
17. Kan MY (2004) Web page categorization without the web page. In: Proceedings of the 13th international World Wide Web conference on Alternate track papers \& posters (WWW Alt.'04). ACM, New York, NY, USA, pp. 262–263
18. Kruse P, Naujoks A, Roesner D, Kunze M (2005) Clever search: a WordNet based wrapper for internet search engines, computing research repository, vol. abs/cs/050
19. Langer L, Frøkjaer E (2008) Improving web search transparency by using a Venn diagram interface. In: Proceedings of the 5th Nordic conference on Human-computer interaction: building bridges (NordiCHI'08). ACM, New York, NY, USA, pp. 249–256
20. Lew MS, Sebe N, Djeraba C, Jain R (2006) Content-based multimedia information retrieval: state of the art and challenges. ACM Trans Multimed Comput Comm Appl 2(1):1–19
21. Lin X (1992) Visualization for the document space. In: Proceedings of the 3rd conference on Visualization'92 (VIS'92). IEEE Computer Society Press, Los Alamitos, CA, USA, pp. 274–281
22. Marchionini G (2006) Exploratory search: from finding to understanding. Comm ACM 49(4):41–46
23. Marcus A, Gould EW (2000) Crosscurrents: cultural dimensions and global web user-interface design. Interactions 7(4):32–46
24. Mukherjea S, Cho J (1999) Automatically determining semantics for world wide web multimedia information retrieval. J Vis Lang Comput 10(6):585–606
25. Nowell LT, France RK, Hix D, Heath LS, Fox EA (1996) Visualizing search results: some alternatives to query-document similarity. In: Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'96). ACM, New York, NY, USA, pp. 67–75
26. Park H, Myaeng S-H, Jang G, Choi J, Jo S, Roh H-C (2008) SketchBrain: an interactive information seeking interface for exploratory search. In: Proceedings International Workshop on Human-Computer Interaction and Information Retrieval (HCIR 2008), pp. 53–56
27. Perer A, Shneiderman B (2009) Integrating statistics and visualization for exploratory power: from long-term case studies to design guidelines. IEEE Comput Graph Appl 29(3):39–51
28. Pirolli P, Card S (1999) Information foraging. Psychol Rev 106(4):643–675
29. Plaisant C, Milash B, Rose A, Widoff S, Shneiderman B (1996) LifeLines: visualizing personal histories. In: Proceedings of the SIGCHI conference on Human factors in computing systems: common ground (CHI'96), ACM, New York, NY, USA, 221–227
30. Porter MF (1980) An algorithm for suffix stripping. Program 14(3):130–137
31. Radev D, Fan W (2000) Automatic summarization of search engine hit lists. In: Proceedings of the ACL-2000 workshop on Recent advances in natural language processing and information retrieval: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics–Volume 11 (RANLPIR'00), vol. 11. Association for Computational Linguistics, Stroudsburg, PA, USA, 99–109
32. Resnik M, Maldonado C, Santos J, Lergier R (2001) Modeling online search behavior using alternative output structures. In: Proceedings of the Human Factors and Ergonomics Society 45th Annual Conference, Minneapolis, MN, USA, pp. 1166–1171
33. Roberts J, Boukhelifa N, Rogers P (2002) Multiform glyph based web search result visualization. In: Proceedings of the 6th IEEE International Conference on Information Visualization, 2002, IEEE Press. pp. 549–554
34. Rose E, Levinson D (2004) Understanding user goals in web search. In: Proceedings of the 13th international conference on World Wide Web (WWW'04). ACM, New York, NY, USA, 13–19
35. Salmeron L, Gil L, Braten I, Stømsø H (2010) Comprehension effects of signalling relationships between documents in search engines. Comput Hum Behav 26(3):419–426
36. Salton G, Buckley C (1988) Term weighting approaches in automatic text retrieval. Inf Process Manag 24(5):513–523, August 1988
37. Santini S, Gupta A, Jain R (2001) Emergent semantics through interaction in image databases. IEEE Trans Knowl Data Eng 13(3):337–351

38. Singh R, Hsu YW (2007) Analysis of usage patterns in experiential multiple perspective web-search. In: Proceedings of the 15th international conference on Multimedia (MULTIMEDIA'07). ACM, New York, NY, USA, 569–572
39. Singh R, Hsu YA-W, Sun WC, Chiture D, Yan L (2005) Rethinking the presentation of results from web search. In: Proceedings of the IEEE International Conference on Multimedia and Expo. pp. 1492–1495
40. Singh R, Jain R (2006) From information centric to experiential environments. In: Goldin D, Smolka S, Wegner P (eds) Interactive computation: the new paradigm. Springer Verlag, pp. 323–351. ISBN: 978–3540346661
41. Singh R, Li Z, Kim P, Pack D, Jain R (2004) Event-based modeling and processing of digital media. In: Proceedings of the 1st international workshop on Computer vision meets databases (CVDB'04). ACM, New York, NY, USA, pp. 19–26
42. Singh R, Pinzon JC (2007) Study and analysis of user behavior and usage patterns in a unified personal multimedia information environment. In: Proceedings of the IEEE International Conference on Multimedia and Expo, pp. 1031–1034, 2007
43. Smeulders AWM, Worring M, Santini S, Gupta A, Jain R (2000) Content based image retrieval at the end of the early years. IEEE Trans Pattern Anal Mach Intell 22(12):1349–1380
44. Snoek CGM, Worring M (2009) Concept-based video retrieval. Found Trends Inform Retrieval 2 (4):215–322
45. Strehl A (2002) Relationship-based clustering and cluster ensembles for high-dimensional data mining, PhD Dissertation, Department of Electrical and Computer Engineering, University of Texas at Austin
46. Teevan J, Alvarado C, Ackerman MS, Karger DR (2004) The perfect search engine is not enough: a study of orienteering behavior in directed search. In: Proceedings of the SIGCHI conference on Human factors in computing systems (CHI'04). ACM, New York, NY, USA, 415–422
47. Toyama K, Logan R, Roseway A, Anandan P (2003) Geographic location tags on digital images. In: Proceedings of the eleventh ACM international conference on Multimedia (MULTIMEDIA'03). ACM, New York, NY, USA, 156–166
48. Vessey I, Galletta D (1991) Cognitive fit: an empirical study of information acquisition. Inf Syst Res 2:63–84
49. Wang X-J, Ma W-Y, He Q-C, Li X (2004) Grouping web image search result. In: Proceedings of the 12th annual ACM international conference on Multimedia (MULTIMEDIA'04). ACM, New York, NY, USA, 436–439
50. White RW, Drucker SM, Marchionini G, Hearst M, Schraefel MC (2007) Exploratory search and HCI: designing and evaluating interfaces to support exploratory search interaction. In: Proceedings of CHI'07 extended abstracts on Human factors in computing systems (CHI EA'07). ACM, New York, NY, USA, pp. 2877–2880
51. White RW, Roth RA (2009) Exploratory search: beyond the query-response paradigm. Morgan and Claypool, San Rafael
52. Wilson ML, Kules B, Schraefel MC, Shneiderman B (2010) From keyword search to exploration: designing future search interfaces for the web. Found Trends Web Sci 2(1):1–97
53. Wise JA, Thomas JJ, Pennock K, Lantrip D, Pottier M, Schur A, Crow V (1999) Visualizing the non-visual: spatial analysis and interaction with information from text documents. In: Card SK, Mackinlay JD, Shneiderman B (eds) Reading in information visualization: using vision to think. Morgan Kaufmann, San Francisco, pp 442–445
54. Woodruff A, Faulring A, Rosenholtz R, Morrsion J, Pirolli P (2001) Using thumbnails to search the Web. In: Proceedings of the SIGCHI conference on Human factors in computing systems (CHI'01). ACM, New York, NY, USA, 198–205
55. Yee K-P, Swearingen K, Li K, Hearst M (2003) Faceted metadata for image search and browsing. In: Proceedings of the SIGCHI conference on Human factors in computing systems (CHI'03). ACM, New York, NY, USA, 401–408
56. Zamir O, Etzioni O (1999) Grouper: a dynamic clustering interface to Web search results. In: Enslow PH Jr (ed) Proceedings of the eighth international conference on World Wide Web (WWW'99). Elsevier North-Holland, Inc., New York, pp 1361–1374
57. Zeng H-J, He Q-C, Chen Z, Ma W-Y, Ma J (2004) Learning to cluster web-search results. In: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'04). ACM, New York, NY, USA, 210–217
58. Zubizarreta Á, Fuente P, Cantera JM, Arias M, Cabrero J, García G, Llamas C, Vegas J (2008) A georeferencing multistage method for locating geographic context in web search. In: Proceeding of the 17th ACM conference on Information and knowledge management (CIKM'08). ACM, New York, NY, USA, pp. 1485–1486



Rahul Singh is currently an associate professor in the department of Computer Science. He received his PhD in Computer Science from the University of Minnesota and his MS in Computer Science from the Moscow Power Engineering Institute. His technical interests are in bioinformatics, computational drug discovery and multimedia information modeling and management. Prior to joining academia, he was in the industry (in the San Francisco-Bay area) at Scimagix, where he worked on various problems related to multimedia biological information management and at Exelixis, where he founded and headed the computational drug discovery group, which worked on various problems across the genomics-drug discovery spectrum. Dr. Singh is a recipient of the CAREER award of the National Science Foundation and was a San Francisco State University presidential fellow.

Ya-Wen Hsu received her MS degree in Computer Science from the San Francisco State. She is currently employed at eline LLC. Ms Hsu was the recipient of the *Outstanding Culminating Research Award*, Department of Computer Science SFSU 2007 and received the *San Francisco State University Graduate Student Distinguished Achievement Award* in 2007.

Naureen Moon was a graduate student in the BioComputing and Media Research Lab at San Francisco State University.