Designing Interaction Paradigms for Web-Information Search and Retrieval

Ya-Wen Hsu, Naureen Moon, Rahul Singh

Department of Computer Science, San Francisco State University, San Francisco, CA 94132
logoin@sfsu.edu, numoon@sfsu.edu, rsingh@cs.sfsu.edu

Abstract

As the complexity of the information available on the web increases, the role of user-data interaction paradigms is becoming increasingly critical for the success of web information retrieval. Recent years have witnessed significant advances in techniques for indexing and querying web data. However, in the same period, limited advancements have been made in developing paradigms and researching algorithmic issues associated with the design of interfaces for web-search. In this paper, we propose a novel paradigm for enabling multiple-perspective query and interaction in web search. Underlying the proposed metaphor are information and pattern analysis techniques that help determine semantic correlations between web pages, identify and extract information critical for intuitive understanding and hypothesis generation, and support effective and multiple-perspective interactions between users and the data. We provide a comprehensive study on the effectiveness and efficiency of the proposed approach in query-retrieval scenarios involving complex information goals. Our investigations point to the importance of developing novel ways to mediate interactions during web-search and will be useful in the development of the next generation of real-world solutions for web information retrieval.

1. Introduction

The traditional model of information retrieval and presentation on the web is based on users providing key query terms to a search engine and the system providing a list of web sites pertinent to the term(s), typically ranked by relevance. Such an approach is highly effective when the queries are well defined and seek precise information. However, with the evolution of information available on the web as well as in the nature of queries, the traditional model is fundamentally challenged to support efficacious information assimilation. For instance, the limitations of the current paradigm can be seen in its handling of polysemous information and information expressed through non-textual media. Similarly, explicitly utilizing spatial and temporal characteristics of information in generic web querying remain complex, even though the importance of location and time in information retrieval and organization is well established [7, 8, 14]. Furthermore, the success of retrieving information in the current paradigm often depends on the selection of appropriate keywords. This emphasizes the unsuitability of the current paradigm for exploratory querying and information assimilation as opposed to information lookup.

Our research seeks to address many of the aforementioned issues by developing a new paradigm for user-data interaction in web-search. In the proposed approach, users can 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. Studies and evaluations presented in this paper indicate that such an approach can overcome several limitations of the current paradigm and be especially helpful in situations such as exploratory information search, finding patterns in the data, understanding the information space induced by the query, and recognizing the information diversity underlying the query. Unlike the traditional approach, where a user is limited to keywords in issuing or refining a query, the proposed approach allows examination of the retrieved information structure from multiple semantic "viewpoints". The proposed paradigm is also interactive, in that, information explorations conducted with respect to a chosen semantic viewpoint (such as, for instance, the temporal characteristics of the information) are reflected in the other semantic viewpoints supported by the data. Thus users can form a holistic understanding of the information space. Furthermore, the proposed paradigm leads to more rapid localization and retrieval of relevant information, when compared to the current approach which is based on ranking and listing the results. Thus, the capabilities supported through this paradigm do not impair the efficiency of user-data interactions, rather they often improve it. Finally, the proposed interaction paradigm is designed to utilize any of the commonly used search engines. This not only allows the paradigm to be used as an interface to the many excellent search engines available today but also obviates the necessity to support separate indexing of the extraordinary volume of data on the web.

2. Previous work

The use of linguistic techniques to analyze search results and infer categories based on content has been the predominant thrust of various approaches for alternate
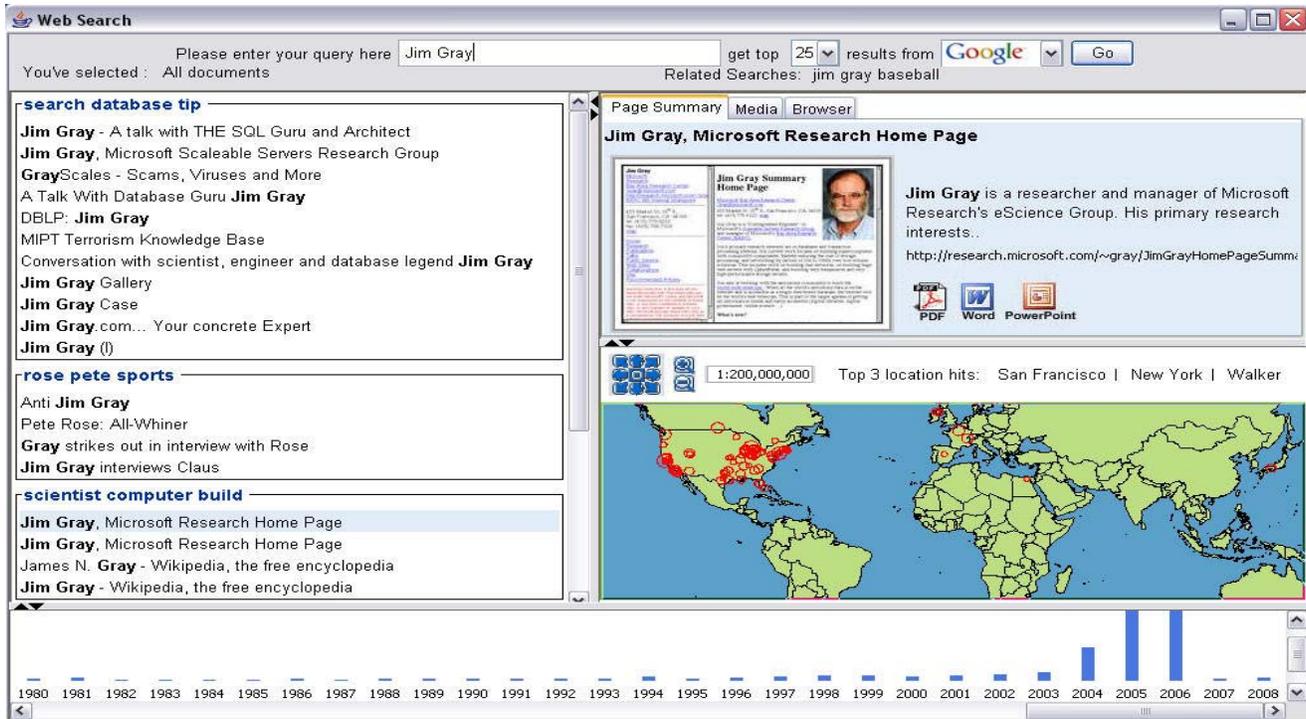
Figure 1. The user interface containing labeled document clustering, page summary, map, and timeline.
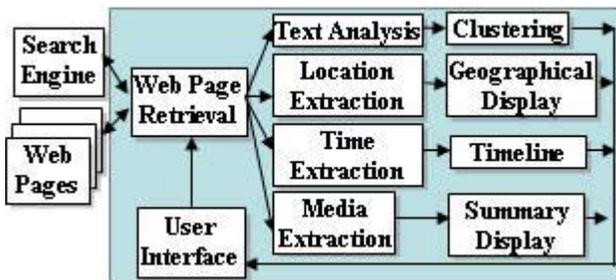

Figure 2. The system architecture.

presentation of web search results. A recent review is povided in [10]. At the state of the art, commercially available solutions to this problem include Vivisimo [16] and Grokker [4]. The Vivisimo search engine provides a text-based clustering capability using a heuristic that looks for subsets having concise descriptions. Grokker on the other hand provides interactive graphical clustering of results. In other research, [3] proposes position sensitive word-based and TFIDF-based clustering while [13] filters search results using WordNet [17] senses. Unlike these approaches, where web page categorization is defined by content, the search engine Northern Light [5] attempts to organize results into categories predefined in library sciences. In [6], the clustering problem is reformulated as a problem of ranking salient phrases, which is then handled in a supervised manner.

Notwithstanding certain technical overlaps, our interpretation of the problem and solution philosophy differs fundamentally. Most of the above approaches consider the problem as that of summarizing hits from

web search. In contrast, our goal is to support data context, provide information overview, and allow proactive querying from different semantically meaningful perspectives. Thus summarization is but one consequence of our approach where aspects such as interactivity and presentation of the data from different semantically-correlated perspectives play an equally important role. In contrast to techniques based on linguistic repositories like WordNet, the main components of our approach are data-driven. This allows us to robustly deal with situations where the query terms and the appropriate sense (such as apple as in computers) are not present in WordNet.

3. Architecture and system overview

A snapshot of the proposed system is shown in Figure 1. The key modules involved in the system and the information flow between them is shown in Figure 2. Queries issued in the system are directed to a search engine of choice, with the *Web Page Retrieval* module obtaining the textual content of the hits. This module also uses Alexa [1] to obtain web page thumbnails. The text-based information is then successively processed by the *Text Analysis*, *Location Extraction*, 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 obtain its temporal information. The design of these modules and the underlying algorithms are described in detail in Sections 4, 5, and 6 respectively.

The information derived from the various components is displayed as separate panes in the user interface window (Figure 1), 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 pane 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. 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 is essential for discovering relations hidden in the data and in formulation of insights.

4. Determining 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 inherent variability in the information present in the corresponding web-pages.

As a preprocessing step towards this goal, the text of each web page is appended to its metadata and parsed to remove all html tags, numbers, punctuation, etc. A stop-list of 670 common English words is used to remove minimally-descriptive terms (e.g., "that") and only the first 100 remaining words are retained. The Porter stemming algorithm [9] is then 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 elements of Latent Semantic Indexing (LSA) with a Term Frequency-Inverse Document Frequency (TFIDF) strategy to enhance the quality of the final document clusters by inducing the greatest correlation between the term frequency vectors of similar documents and producing the greatest distance between dissimilar documents. The intuition lies in utilizing LSA to reveal semantic correlations between the documents via mapping to an eigenspace, but then to augment the prominence of uncommon words through TFIDF.

After normalization (dividing frequency values by document size), LSA is performed on the term frequency table. The technique consists of performing singular value decomposition (SVD) on the term frequency matrix to rewrite it as a product of 3 matrices:

$$\{X\} = \{W\}\{S\}\{P\}^T \quad (1)$$

where $\{S\}$ is a diagonal matrix of scaling values. The dimensionality of the solution is then reduced by truncating all but the largest eigenvectors, in order to minimize the amount of noise in the data set. We automatically select the number of dimensions by calculating the difference between contiguous eigenvalues and truncating all values after the largest drop. Lastly, the matrix is reconstructed to reveal the modified term frequency values.

We then further adjust the values using TFIDF, which 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 follows (where the latter term is called the inverse document frequency):

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

where $TF(i,j)$ is the frequency of term i in document j, N the size of background set, and $DF(i)$ the number of background set documents in which i appears.

Since the performance of TFIDF is heavily influenced by the choice of background set, we construct a large background set of 1656 random web pages and compare its performance with that of a small, dynamically-generated background set consisting of the first 50 search results. The latter is motivated by the observation that the discernment value of a term depends upon context. For example, the term "quark" is uncommon in normal English usage, but would occur frequently in the results for a query on particle physics so that its ability to distinguish between documents in that domain would be very limited. The results (Figure 3) show that the dynamic background set does indeed yield considerably better and more consistent values. Finally, the correlation is measured using the Pearson's coefficient, which was empirically determined to perform slightly better than the cosine similarity measure and considerably better than Euclidean similarity.
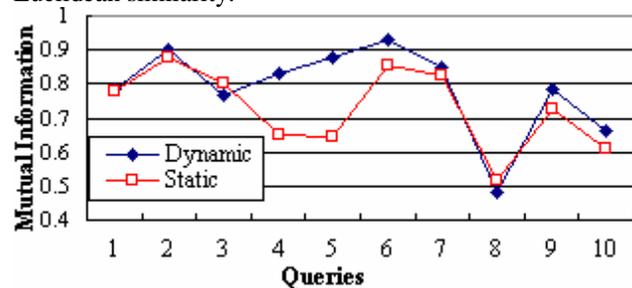

Figure 3. Comparison of background sets for TFIDF.

5. Data and Model-Based Clustering of Correlated Pages

We present a two-stage clustering mechanism for document correlation. The first pass (to which the above discussion pertains) is dynamic and data-driven while the second is model-based and uses taxonomy information (when available) to refine the results. This is a significant addition because it allows similar documents to be brought together despite a possible lack of correlating data. For example, two web pages about airlines might employ very different terms to describe their services. In this case, the data-driven clustering may assign a low correlation value to the documents, but to humans the semantic relationship is quite clear and may therefore be captured by manually-constructed taxonomies.

5.1. Date-driven clustering

To implement data-driven clustering, we designed an adaptive K-medoid clustering algorithm. This algorithm is designed to address two major issues associated with the K-means family of algorithms:
- Specifying K, the number of clusters, in advance.
- Ameliorating the dependence of the final result on the cluster initialization.

To solve the first problem, we use an adaptive, data-driven approach to determine K; we identify the K-highest density points in the distribution to seed the clusters. Additionally, we employ the K-medoids algorithm, which uses actual data points as centers of clusters and therefore requires less re-computation of the distances between objects and cluster medoids since the medoids change less often as compared to cluster means.

If any two objects in a given cluster have similarity value lower than a threshold T (empirically selected to be 0.93) , the cluster becomes a candidate for "splitting". Analogously, for two clusters, if all documents have pair-wise similarity higher than T, the clusters become candidates for "merging". A cluster is defined to be "stable" if for any two documents in the cluster, the similarity value is higher than T while that of a document in this cluster with any document in another cluster is less than T. The point at which all clusters become stable determines K and acts as a stopping criterion.

5.2. Model-based clustering

ODP [11] 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"). We first obtain the categorization(s), when available, for each retrieved web page, truncating the last level for web pages with a hierarchy depth of 8 or more levels. Then we merge clusters containing documents with either the same categorization, or with parent-child categorization since the parent is merely a more generalized instance of the child.

Clusters are labeled using the three keywords with the highest document frequency within the cluster. 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."

Figure 4. Selection menu in location display module.

6. Interactions with location and time information

The full text from each web page is parsed to extract location information through cross-referencing with an index of locations constructed from World Gazetteer [18]. Location name ambiguity is dealt with by looking for a city name only if the country or state of which it is a part is found in the document text (with exceptions made for major world cities).

Display and interaction is implemented through the use of the OpenMap Java toolkit [12]. Cities appearing in the documents are indicated on the map, where circle size varies logarithmically with the number of documents containing the location. The user may interact with the spatial information using both point-based and region-based queries (through clicking on the area or dragging a rectangle). To disambiguate user intent, a list of locations contained within the selection is displayed and the user may select one or all of them (Figure 4).

The document texts are also parsed to extract their temporal information which is displayed through a timeline that supports multiple granularities so that users may expand year view to month and day views.

The media, map, and timeline modalities all enable visual overviews of the information. These views support direct manipulations and are reflective; selecting information in any single pane automatically highlights its corresponding characteristics in the others. Users can thus interact and explore the information using any of the

available views (images, text-based semantic similarity, spatial characteristics, and/or temporal characteristics).

7. System Evaluation

7.1 Analysis of Clustering

We note that the purposes of clustering documents include: (1) reduction of the cognitive load for the user by decreasing the number of links to peruse and (2) Providing an understanding of the information space by bringing related links/motifs together. To assess performance with respect to the first criteria, we count the number of (non-singleton) clusters as an indicator of the reduction in the number of information cues a user needs to pursue. Success in attaining the second objective is determined by appraising the correctness of clusters using the mutual information measure, which quantifies the agreement between algorithm-derived and ground-truth clusters, which are manually defined. Unlike other cluster quality measures, mutual information does not favor small clusters. We use the [0,1]-normalized mutual information measure [2], where a value of 1 indicates perfect clustering while random clustering gives a value of 0 in the limit.

For the purposes of this experiment we consider 25 queries and analyze the top 20 documents retrieved. A dynamic background set of 50 documents is used with an empirically-determined first-stage clustering threshold of 0.90. Results are generated for TFIDF on LSA both with and without category-based clustering to quantify the advantages (if any) is obtained through the addition of a second clustering stage.

It is noteworthy 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, large clusters would lead to large data reduction, but low cluster coherence and therefore low mutual information values. This indicates that data reduction and mutual information values must be viewed jointly.

The results for the first experiment (Table 1) show an average data size reduction of 60% with categories and 51% without, which is calculated as the percent difference between the number of clusters and the number of documents they contain. This level of decrease in data volume is clearly advantageous for the user.

We also evaluate the improvement in clustering produced by adding TFIDF to LSA, and that caused by adding category-based clustering. The results (Figure 5) show that the addition of TFIDF does indeed improve the mutual information of the final clusters in most cases, on average by about 0.03. This verifies that the increased in-cluster and out-of-cluster difference does indeed translate

to a better final clustering result. Five queries (4, 6, 16, 21, and 22) show decreases in mutual information after the addition of TFIDF (see Figure 5). All of these include results that are noisy (containing semantically meaningless information such as addresses) and/or lack text. It is difficult to correctly cluster the documents in this case, and the difference is augmented by TFIDF. However, both with and without TFIDF, the mutual information of all queries is above 0.5.
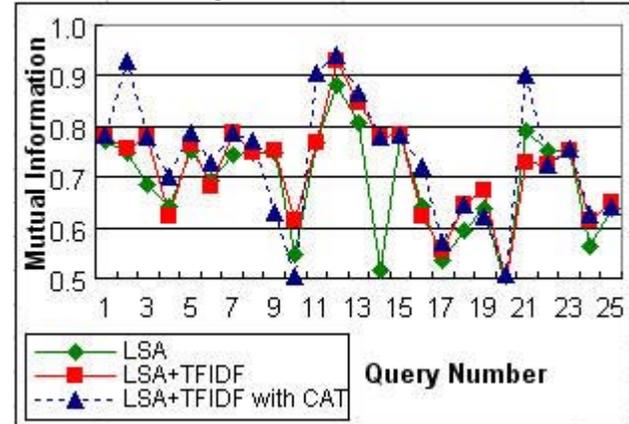

Figure 5. **Comparison of different analysis approaches for grouping semantically-related pages.**

Table 1. Data reduction due to clustering.

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
	Average	10	51	8	60

The additional clustering stage shows better or equally good results in the vast majority of cases. The exceptions include cases like the 9th, 10th, 19th and 25th queries.

This is typically due to one of two reasons: (1) one or more of the original clusters are inconsistent, so their merger is even more so or (2) the original clusters are coherent, but it is inappropriate to merge them despite the category match. The effect of category-based clustering may also be limited due to (1) absence of category information, (2) lack of category matches, or (3) documents with the same categories already mapped to the same cluster by data-driven clustering.

While we do find that the combination of LSA and category-based clustering significantly help to capture the semantics of the information, it is not a panacea. For instance, it is possible to find cases where noise and lack of data lead to problems in analysis. In such cases, variations in system parameters and/or inclusion of additional information, such as meta-data about the page have been observed to be of help. Another possibility would be to infer the page content from the content of other pages hyperlinked to it.

7.2. Robustness of location and time extraction

The effectiveness of the location and time extraction modules was evaluated by comparison with manual extraction. The test set consisted of the first 20 results for each of 3 queries ("eclipse," "computer," and "cell phone"). Both modules show values in the mid-90% range for recall, indicating that the vast majority of geographical and temporal information is extracted. As for precision, 99% of extracted temporal data and about 88% of extracted geographical data are correct. Precision and recall values for time extraction are as expected because the algorithm detects the vast majority of date formats and is rarely incorrect. The comparatively low value for location precision is due primarily to the aforementioned name ambiguity.

8. User Evaluation

The purpose of the study was to evaluate how well the proposed paradigm helps participants on searching and exploring information compare to some of the most commonly used commercial systems. In particular, we focused on how participants utilize each of the modalities in our system and whether our system is considered more useful to locate desired information after a period of use.

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 second at the beginning of next week. In between the two sections, participants were expected to continue to use our system in order to further familiarize themselves and gain expertise with different features in the system. Two question sets were interchangeably used to eliminate any bias associated with ordering. The purpose of the

extended study was to estimate any improvement in how successfully participants use the various systems, with prolonged use.

Each questionnaire included 20 tasks (queries); each task required participants to issue a query and target desired information on designated system(s). Of these, 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. While space constraints prevent us from exhaustively listing these queries, in the following we list some of them and point out in brackets, the primary information modality/system module relevant to the query: (Location-oriented): "where can pandas be found in the US?", "Asian country where mummies were discovered", (Time-oriented): "When did the store IKEA open in Taipei", "Which year was LINUX released?", (Media-oriented): "Find an audio about the recent immigration-protests in the US". "Find a video about Heaven's gate", (Text-oriented): "List the five hottest varieties of chili-pepper", "Find uses of Lavender in cooking and medicinal therapy". The results of the queries were collected and used in evaluating the user responses. As may be observed, many of these queries have characteristics that can be explored using more than one semantic perspective, such by combining space and time.

Twenty participants were recruited and 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 is important since participants have a sense of the information content pertinent to a query after they have run it on the first system. Therefore, without such precautions, the results would favor the last three applications being evaluated.

Participants were allocated a period of 3 minutes (excluding system retrieval time) to perform a task, after which they had to proceed to the next task. In order to measure the efficiency of information access, the time and the number of clicks associated with each task were recorded. At the end of the study, participants were asked to provide a qualitative evaluation (using a scale of 1 to 5) of all the four applications in terms of how well they facilitated information search and exploration.

8.1. Results and Analysis

The data from the user studies was used to test the following hypotheses:
1. Participants will experience greater search satisfaction and success with our system than with other commercial search engines.
2. Compared with Vivisimo, our system can give participants a better insight and informational structure of a query in a shorter time.

3. Participants will gradually learn to use each of the components and obtain desired information with less time.

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 is less than 0.05, two variables are considered significant different. Between groups of estimate of the variable variance is much larger than the within group estimate when F value greater than 1. We use pair-wise comparison of different variables to analyze the affect of different factors involved the study.

8.1.1. Comparison with commercial system: For every query, we compare the time and number of clicks needed to find the desired information with respect to Google, Vivisimo, and Grokker. As shown in Figure 6 and Figure 7, our system has a noticeably better result on both measurements. That is, participants were able to reach information goal with at least 20 seconds faster or 6 fewer clicks than all 3 commercial systems. The main effect of participants on the needed time to reach desired information was significant for Google ($F=6.43$, $p \ll 0.001$), Vivisimo ($F=4.35$, $p \ll 0.001$) and Grokker ($F=2.425$, $p \ll 0.001$). Even though our system shows variability between participants ($F=1.82$, $p=0.018$), all participants were able to reach their information goal with less time using our application, which displays considerably lower values for the standard deviation compared to the others (the p-value of our system is greater than others). On the contrary, the affect of participants is not significant in terms of number of clicks ($p \ll 0.004$ in all applications). Research in [15] shows similar variability between search time for users, and uses log time to normalize values.

8.1.2. 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 possibly 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, this corresponds 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 study it using six queries and compare it with Vivisimo, where the notion of such semantic grouping of results is also supported. As part of this experiment, participants were

asked to explore information relevant to six information goals using the proposed system and Vivisimo. At the end, the participants rated the systems on a 1 to 5 scale where 1 is very difficult and 5 very easy. The results from the experiment are shown in Figure 8.
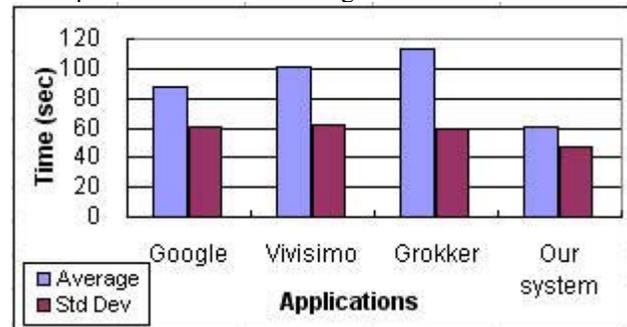

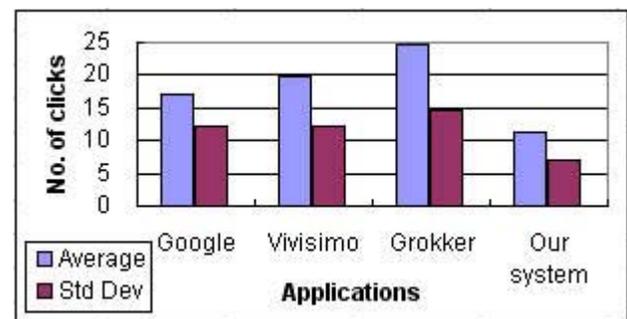**Figure 6. Average time to reach information goal and its standard deviation.**

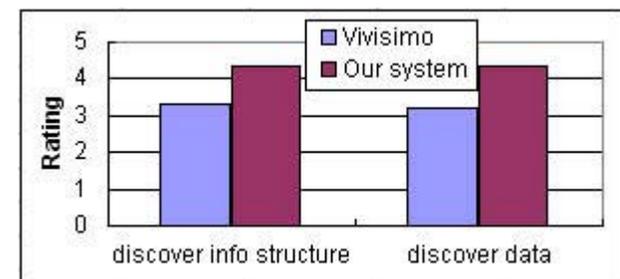**Figure 7. Average no. of clicks to reach information goal and its standard deviation.**

Figure 8. Subjective evaluation of the information discovery capabilities (1 is very difficult and 5 very easy). Average score for proposed system 4.33

8.1.3. Analysis of the usage of different components: Given the various perspectives on the information supported in our approach, an interesting question arises as to how important users find each module as they search for specific information goals. To evaluate this, user responses on the 28 queries used as part of the user study (14 queries per session) were analyzed. It is interesting to note that from our perspective as study designers the necessary information relevant to 36% of the queries could be retrieved using the text-clustering, 21% using

spatial display, 18% using temporal display, and 25% using media display.

At the conclusion of each query, each of the 20 users 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. The responses were text-clustering (47%), spatial-display (19%), temporal-display (12%), and media-display (22%). The choice of spatial-clustering for a greater number of problems (than what was anticipated by us) is noteworthy. We postulate that this may be due to the fact that text-based interactions have historically been the predominant way of web-information retrieval. Thus users were more inclined to use textual modality to find the necessary information. It is also intuitive to note, in this context, that queries regarding media, took the least time to answer for most users.

Finally, we consider 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 seconds on Google, 105 seconds on Vivisimo, 114 seconds on Grokker and 66 seconds 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. The fact that the proposed approach required less time-to-information amongst all the systems in both stages, illustrates the applicability of alternate paradigms to mediate user-data interactions in web search, especially for complex exploratory queries.

9. Conclusion

In this paper, we have presented research that re-thinks the problem of web-search from the perspective of designing novel user-data interaction paradigms. As part of this investigation, we considered two major design tasks: (1) determination and effective presentation of semantic correlations between web pages and (2) design of a unified presentation-query-exploration interface that facilitates direct interactions between the users and the data. Evaluation of the proposed approach show promising results in terms of its applicability for complex information retrieval and assimilation tasks on the web.

We therefore believe that multi-perspective and interactive front ends for web-search, such as the one presented in this paper, hold significant promise in the design of the next generation of information retrieval systems targeting the web.

10. Author Contributions

Research formulation and overall solution methodology (RS), system design (RS and YWH), primary implementation and overall integration (YWH), location and time extraction (NM), text analysis algorithms: design (RS), and implementation (YWH), experiment design (RS, YWH, NM), data collection and user trials (YWH, NM), data analysis (YWH). This paper was written by RS with contributions from Y-WH and NM. The authors would also like to acknowledge Liu Yan and Dil Chitaure who worked with RS in designing the clustering approach and map interface respectively.

11. References

[1] Alexa, http://www.alexa.com.
[2] A. Strehl, "Relationship-Based Clustering and Cluster Ensembles for High-Dimensional Data Mining," PhD Diss., UT Austin, 2002.
[3] D. Radev and W. Fan, "Automatic Summarization of Search Engine Hit Lists," ACL Workshop on Recent Advancements in NLP and IR, Hong Kong, 2000.
[4] Grokker, http://www.grokker.com.
[5] G. Notess, "Review of Northern Light," 2002.
[6] H-J Zeng, Q-C He, Z. Chen, W-Y Ma, and J Ma, "Learning to Cluster Web-Search Results", SIGIR, United Kingdom, July 2004, p.210-217.
[7] J. Rekimoto, "Time-machine computing: a time-centric approach for the information environment", ACM Symposium on User Interface Software and Technology, 1999, p.45-54.
[8] K. Toyama et al, "Geographic Location Tags on Digital Images", ACM Multimedia, 2003, p.156-166.
[9] M. F. Porter,"An algorithm for suffix stripping," Program, 14(3), 1980, p.130-137.
[10] M. Hicks, "Search Startups Target Clustering," eWeek.com, 2004.
[11] Open Directory Project, http://www.dmoz.org.
[12] OpenMap, http://openmap.bbn.com.
[13] P. Kruse et al, "Clever Search: A WordNet Based Wrapper for Internet Search Engines," Proc. 2nd GermaNet Workshop, 2005.
[14] R. Singh et al, "Designing Experiential Environments for Management of Personal Multimedia," ACM Multimedia, 2004, p.496-499.
[15] S. Dumais, E. Cutrell, and H. Chen, "Optimizing search by showing results in context," ACM SIGCHI, 2001, p.277-284.
[16] Vivisimo, http://www.vivisimo.com.
[17] WordNet, http://wordnet.princeton.edu.
[18] World Gazetteer, http://www.world-gazetteer.com.