PERSONALIZED AND COLLABORATIVE CLUSTERING OF SEARCH RESULTS

THESIS

Presented to the Graduate Council
of Texas State University-San Marcos
in Partial Fulfillment
of the Requirements

for the Degree

Master of SCIENCE

by

Dragos Anastasiu, B.A.

San Marcos, Texas
August 2011

PERSONALIZED AND COLLABORATIVE CLUSTERING OF SEARCH RESULTS

Committee Members Approved:

Byron J. Gao, Chair

Anne H.H. Ngu

Yijuan Lu

Approved:

J. Michael Willoughby
Dean of the Graduate College

FAIR USE AND AUTHOR'S PERMISSION STATEMENT

Fair Use

Duplication Permission

ACKNOWLEDGEMENTS

Many thanks go to Dr. Byron J. Gao, without whose guidance and inspiration I would not be involved in research. I would like to thank Dr. Anne H.H. Ngu and Dr. Yijuan Lu for their constant support and encouragement during the thesis process. I would also like to thank Dr. David Buttler from Lawrence Livermore National Laboratory, who gave me invaluable conceptual and implementation insight and feedback. Finally, I wish to thank my family for their patience and understanding during my many late nights at the office the past year or two.

This manuscript was submitted on June 22, 2011.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

PERSONALIZED AND COLLABORATIVE CLUSTERING OF SEARCH RESULTS

by

Dragos Anastasiu, B.A.

Texas State University-San Marcos

August 2011

SUPERVISING PROFESSOR: BYRON J. GAO

Organizing and presenting search results plays a critical role in the utility of search engines. Due to the unprecedented scale of the Web and diversity of search results, the common strategy of ranked lists has become increasingly inadequate, and clustering has been considered as a promising alternative. Clustering divides a long list of disparate search results into a few topic-coherent clusters, allowing the user to quickly locate relevant results by topic navigation. While many clustering algorithms have been proposed that innovate on the automatic clustering procedure, I introduce ClusteringWiki, the first prototype and framework for personalized clustering that allows direct user editing of the clustering results. Through a Wiki

interface, the user can edit and annotate the membership, structure and labels of clusters for a personalized presentation. In addition, the edits and annotations can be shared among users as a mass-collaborative way of improving search result organization and search engine utility.

CHAPTER I

INTRODUCTION

We live in the information age. Billions of documents on all topics imaginable are connected and accessible on the Web through the simple concept of the hyperlink. Yet the sheer size of the Web makes browsing to locate desired information a daunting task. Web search attempts to alleviate this problem by connecting short phrase queries to relevant documents on the Web, which are generally displayed in a flat ranked list.

Every day millions of people search the Web, unaware of the complexity involved in matching their query with the information they seek. They hope that the exact search results they are looking for will be displayed as soon as they execute their query. However, queries are inherently ambiguous and search results are often diverse with multiple senses. With a list presentation, the results on different sub-topics of a query will be mixed together. The user has to sift through many irrelevant results to locate those relevant ones.

With the rapid growth in the scale of the Web, queries have become more ambiguous than ever. For example, there are more than 20 entries in Wikipedia for different renown individuals under the name of Jim Gray, including a computer scientist, a diplomat, a linguist, a poet, a turbine design engineer, a filmmaker, and so on. Suppose we intend to find information about `Jim Gray,` the Turing Award winner, we can issue a query of “Jim Gray” in Yahoo![1]. For this extremely famous name in computer science, only 2 are relevant in the top 10 results.

The way search results are organized and presented has a direct and significant impact on the utility of search engines. While the flat ranked list presentation is acceptable for homogeneous search results, the diversity of search results for most queries has increased to the point that we must consider alternative presentations by providing additional structure to flat lists so as to effectively minimize browsing effort and alleviate information overload [Carpineto et al., 2009; Hearst and Pedersen, 1996; Pirolli et al., 1996; Zamir and Etzioni, 1998]. Over the years clustering has been accepted as the most promising alternative.

Clustering is the process of organizing objects into groups or clusters that exhibit internal cohesion and external isolation. Based on the common observation that it is much easier to scan a few topic-coherent groups than many individual documents, clustering can be used to categorize a long list of disparate search results into a few clusters such that each cluster represents a homogeneous sub-topic of the query. Meaningfully labeled, these clusters form a topic-wise non-predefined,

---

[1] Other choices of search engine in this example would not change the validity of the observations. Also note that search results and their ranks may change over time.

faceted search interface, allowing the user to quickly locate relevant and interesting results. Evidence shows that clustering improves user experience and search result quality [Manning et al., 2008].

Given the significant potential benefits, search result clustering has received increasing attention in recent years from the communities of information retrieval (IR), Web search, and data mining. Many clustering algorithms have been proposed [Hearst and Pedersen, 1996; Kummamuru et al., 2004; Lee et al., 2009; Pirolli et al., 1996; Wang and Zhai, 2007; Zamir and Etzioni, 1998, 1999; Zeng et al., 2004]. In the industry, well-known cluster-based commercial search engines include Clusty (www.clusty.com), iBoogie (www.iboogie.com) and CarrotSearch (carrotsearch.com). Carrot2 (www.carrot2.org) is an open source clustering engine distributed under the BSD license.

Despite the high promise of the approach and a decade of endeavor, cluster-based search engines have not gained prominent popularity, evident by Clusty's Alexa rank [Iskold, 2007]. This is because clustering is known to be a hard problem, and search result clustering is particularly hard due to its high dimensionality, complex semantics and unique additional requirements beyond traditional clustering.

As emphasized in [Wang and Zhai, 2007] and [Carpineto et al., 2009], the primary focus of search result clustering is NOT to produce optimal clusters, an objective that has been pursued for decades for traditional clustering with many successful automatic algorithms. Search result clustering is a highly user-centric task with *two unique additional requirements*. First, clusters must form interesting sub-topics or facets from the user's perspective. Second, clusters must be assigned informative, expressive, meaningful and concise labels. Automatic algorithms often fail to fulfill the human factors in the objectives of search result clustering, generating meaningless, awkward or nonsense cluster labels [Carpineto et al., 2009].

In this thesis, I explore a completely different direction in tackling the problem of clustering search results, utilizing the power of direct user intervention and mass-collaboration. I introduce ClusteringWiki, the first prototype and framework for personalized clustering that allows direct *user* editing of the *clustering results*. This is in sharp contrast with existing approaches that innovate on the *automatic* algorithmic *clustering procedure*.

In ClusteringWiki [Anastasiu et al., 2011], the user can edit and annotate the membership, structure and labels of clusters through a Wiki interface to personalize their search result presentation. Personalization provides direct and immediate benefit to the user by reducing user effort spent locating desired results. Edits and annotations can be implicitly shared among users as a mass-collaborative way of improving search result organization and search engine utility. This approach is in the same spirit as other current trends in the Web, like Web 2.0, semantic web, personalization, social tagging and mass collaboration.

In social tagging, or collaborative tagging, users annotate Web objects, and such personal annotations can be used to collectively classify and find information. ClusteringWiki extends conventional tagging by allowing tagging of structured objects, which are clusters of search results organized in a hierarchy.

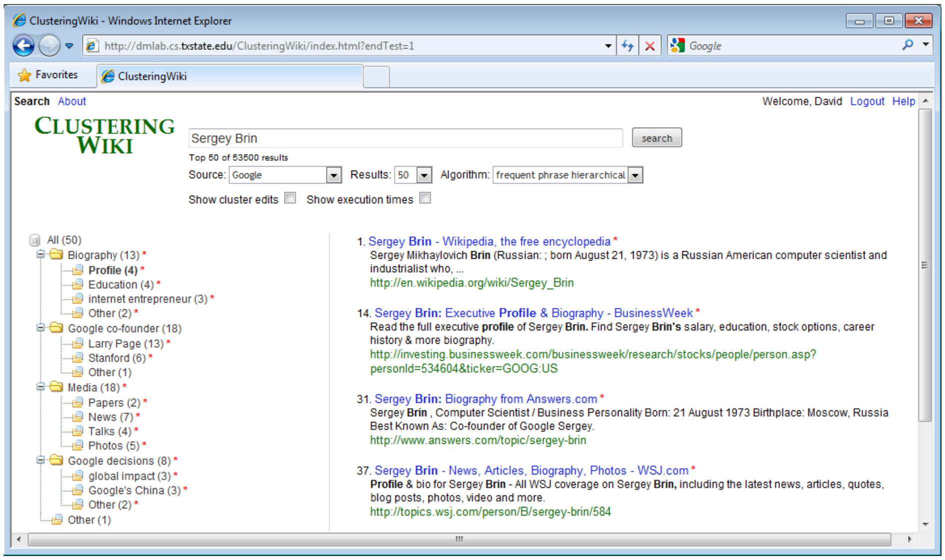
Clustering algorithms fall into two categories: partitioning and hierarchical. Regarding clustering results, however, a hierarchical presentation generalizes a flat partition. Based on this observation, ClusteringWiki handles both clustering methods smoothly by providing editing facilities for cluster hierarchies and treating partitions as a special case. In practice, hierarchical methods are advantageous in clustering search results because they construct a topic hierarchy that allows the user to easily navigate search results at different levels of granularity.

Figure 1.1 shows a snapshot of ClusteringWiki[2]. The left-hand *label panel* presents a hierarchy of cluster labels. The right-hand *result panel* presents search results for a chosen cluster label. A logged-in user can edit the current clusters by creating, deleting, modifying, moving or copying nodes in the cluster tree. Each edit will be validated against a set of predefined consistency constraints before being stored.

Designing and implementing ClusteringWiki pose non-trivial technical

---

[2] dmlab.cs.txstate.edu/ClusteringWiki/.


ClusteringWiki - Windows Internet Explorer
http://dmlab.cs.txstate.edu/ClusteringWiki/index.html?endTest=1
Google
Favorites
ClusteringWiki
Search
About
Welcome, David
Logout
Help
CLUSTERING WIKI
search

Sergey Brin

Top 50 of 53500 results

Source: Google Results: 50 Algorithm: frequent phrase hierarchical

Show cluster edits ☐ Show execution times ☐

- All (50)
  - Biography (13) *
    - **Profile (4)** *
    - Education (4) *
    - internet entrepreneur (3) *
    - Other (2) *
  - Google co-founder (18)
    - Larry Page (13) *
    - Stanford (6) *
    - Other (1)
  - Media (18) *
    - Papers (2) *
    - News (7) *
    - Talks (4) *
    - Photos (5) *
  - Google decisions (8) *
    - global impact (3) *
    - Google's China (3) *
    - Other (2) *
  - Other (1)

1. Sergey **Brin** - Wikipedia, the free encyclopedia *
Sergey Mikhaylovich **Brin** (Russian: ; born August 21, 1973) is a Russian American computer scientist and industrialist who, ...
http://en.wikipedia.org/wiki/Sergey_Brin

14. Sergey **Brin:** Executive **Profile** & Biography - BusinessWeek *
Read the full executive **profile** of Sergey **Brin.** Find Sergey **Brin's** salary, education, stock options, career history & more biography.
http://investing.businessweek.com/businessweek/research/stocks/people/person.asp?personId=534604&ticker=GOOG:US

31. Sergey **Brin:** Biography from Answers.com *
Sergey **Brin** , Computer Scientist / Business Personality Born: 21 August 1973 Birthplace: Moscow, Russia Best Known As: Co-founder of Google Sergey.
http://www.answers.com/topic/sergey-brin

37. Sergey **Brin** - News, Articles, Biography, Photos - WSJ.com *
Profile & bio for Sergey **Brin** - All WSJ coverage on Sergey **Brin,** including the latest news, articles, quotes, blog posts, photos, video and more.
http://topics.wsj.com/person/B/sergey-brin/584

Figure 1.1: Snapshot of ClusteringWiki.

challenges. User edits represent user preferences or constraints that should be respected and enforced when the same query is next issued. Query processing is time-critical, thus efficiency must be given high priority in maintaining and enforcing user preferences. Moreover, complications also come from the dynamic nature of search results that constantly change over time.

Cluster editing takes user effort. It is essential that such user effort can be properly reused. ClusteringWiki considers two kinds of reuse scenarios, *preference transfer* and *preference sharing*. The former transfers user preferences from one query to similar ones, e.g., from “David J. Dewitt” to “David Dewitt.” The latter aggregates and shares clustering preferences among users. Proper aggregation allows users to collaborate at a mass scale and “vote” for the best clustering presentation.

CHAPTER II

BACKGROUND

The World Wide Web was created in 1990 as a result on Sir Tim Barners-Lee's vision for a decentralized system for information dissemination [Zimmerman, 2000]. Since then it has grown exponentially both in terms of number of users and linked documents. Today there are over 17.47 billion estimated pages[1] on the Web, not including documents hidden behind web forms or ftp servers (hidden web documents). This explosion in both the size and depth of the Web makes "browsing" as the main means of finding Web information obsolete.

The research community has been active over the past several decades, investigating new methods of analyzing, organizing, and presenting Web documents, with the goal of minimizing the time spent between executing the user query and filling the information need. Below I present some of the related research which either influences or enables the work in this thesis.

---

[1] Retrieved from www.worldwidewebsize.com on Tuesday, 14 June, 2011

2.1 Web Search

Information retrieval (IR) aims to retrieve, from a large collection, those materials (usually documents) that satisfy an information need [Manning et al., 2008]. When applied to the Web, IR focuses on free-text documents and multimedia files, and is better known as Web search.

[. . .]

The following tables are from another chapter, included here as an example:

Table 2.1: Efficiency evaluation using Yahoo! data source

Number of results	**100**		**200**		**300**		**400**		**500**	
Type of clustering	`F`	`H`	`F`	`H`	`F`	`H`	`F`	`H`	`F`	`H`
Retrieving results	0.979	1.018	1.309	1.222	1.615	1.391	1.847	1.579	1.679	1.661
Preprocessing	0.009	0.011	0.052	0.052	0.037	0.037	0.049	0.112	0.152	0.150
Initial clustering	0.004	0.005	0.051	0.040	0.033	0.042	0.104	0.063	0.118	0.144
Applying preferences	0.006	0.007	0.049	0.012	0.015	0.011	0.021	0.015	0.08	0.013
Presenting final tree	0.143	0.172	0.249	0.278	0.341	0.421	0.451	0.66	0.723	0.752
Other	0.396	0.416	0.469	0.524	0.624	0.684	0.558	0.593	0.853	0.912
Total execution time	0.160	0.194	0.401	0.381	0.426	0.511	0.624	0.850	1.073	1.059
Total response time	1.535	1.628	2.179	2.127	2.665	2.585	3.029	3.022	3.604	3.632

Tables 2.1 and 2.2 show the averaged (over 10 queries) runtime in seconds for all 6 portions of the total response time for the two tested data sources. In addition, I computed and list the average *total execution time*, which includes preprocessing, initial clustering, applying preferences and presenting the final tree. This is the time that the prototype is responsible for. The remaining time is irrelevant to the way

Table 2.2: Efficiency evaluation using New York Times data source

Number of results	**100**		**200**		**300**		**400**		**500**	
Type of clustering	`F`	`H`	`F`	`H`	`F`	`H`	`F`	`H`	`F`	`H`
Retrieving results	0.102	0.113	0.131	0.130	0.285	0.259	0.326	0.338	0.419	0.387
Preprocessing	0.019	0.019	0.067	0.066	0.068	0.069	0.096	0.168	0.189	0.187
Initial clustering	0.005	0.006	0.022	0.025	0.035	0.041	0.053	0.062	0.147	0.196
Applying preferences	0.012	0.011	0.013	0.024	0.018	0.011	0.013	0.017	0.016	0.014
Presenting final tree	0.282	0.279	0.372	0.449	0.591	0.691	0.751	0.872	0.846	0.941
Other	0.338	0.497	0.478	0.678	0.589	0.672	0.625	0.937	0.736	0.868
Total execution time	0.317	0.315	0.473	0.565	0.713	0.813	0.913	1.118	1.197	1.338
Total response time	0.758	0.925	1.083	1.373	1.587	1.744	1.863	2.393	2.352	2.593

the prototype is designed and implemented. From the table we can see that:

- The majority of the total response time is taken up by retrieving search results, which would be negligible if ClusteringWiki was implemented by a search company.
- Applying preferences takes less than 1/10 second in all test cases, which certifies the efficiency of my “path approach” for managing preferences.
- Presenting the final tree takes the majority (roughly 80%) of the total execution time, which can be improved by using alternate user interface technologies.

CHAPTER III

CONCLUSION

Search engine utility has been significantly hampered due to the ever-increasing information overload. Clustering has been considered a promising alternative to ranked lists in improving search result organization. Given the unique human factor in search result clustering, traditional automatic algorithms often fail to generate clusters and labels that are interesting and meaningful from the user's perspective. In this thesis I introduced ClusteringWiki, the first prototype and framework for personalized clustering, utilizing the power of direct user intervention and mass-collaboration. Through a Wiki interface, the user can edit the membership, structure and labels of clusters. Such edits can be aggregated and shared among users to improve search result organization and search engine utility.

Both personalized and collaborative clustering of search results aid users in locating those search results they seek. Personalized clustering saves user effort by allowing the user to place results in familiar clusters. Aggregated clustering also provides significant benefits and is "free," in the sense that it does not take user editing effort.

As an alternate method of personalized and collaborative clustering of search results, I presented ClusteringWiki2, a cluster editing system based on annotations. With complete control over both positive and negative terms and phrases in annotations, users can have the same editing freedom as in ClusteringWiki, while maintaining collaborative transparency.

BIBLIOGRAPHY

Anastasiu, D. C., Buttler, D., and Gao, B. J. (2011). Clusteringwiki: Personalized and collaborative clustering of search results. In *Proceeding of the 34th international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '11, New York, NY, USA. ACM.

Carpineto, C., Osiński, S., Romano, G., and Weiss, D. (2009). A survey of web clustering engines. *ACM Computing Surveys (CSUR)*, 41(3):1–38.

Hearst, M. A. and Pedersen, J. O. (1996). Reexamining the cluster hypothesis: scatter/gather on retrieval results. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '96, pages 76–84, New York, NY, USA. ACM.

Iskold, A. (2007). Overview of clustering and clusty search engine. *www.readwriteweb.com/archives/overview_of_clu.php*.

Kummamuru, K., Lotlikar, R., Roy, S., Singal, K., and Krishnapuram, R. (2004). A hierarchical monothetic document clustering algorithm for summarization and browsing search results. In *Proceedings of the 13th international conference on World Wide Web*, WWW '04, pages 658–665, New York, NY, USA. ACM.

Lee, J., Hwang, S.-w., Nie, Z., and Wen, J.-R. (2009). Query result clustering for object-level search. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 1205–1214, New York, NY, USA. ACM.

Manning, C. D., Raghavan, P., and Schtze, H. (2008). *Introduction to Information Retrieval.* Cambridge University Press.

Pirolli, P., Schank, P., Hearst, M., and Diehl, C. (1996). Scatter/gather browsing communicates the topic structure of a very large text collection. In *Proceedings of the SIGCHI conference on Human factors in computing systems: common ground*, CHI '96, pages 213–220, New York, NY, USA. ACM.

Wang, X. and Zhai, C. (2007). Learn from web search logs to organize search results. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, pages 87–94, New York, NY, USA. ACM.

Zamir, O. and Etzioni, O. (1998). Web document clustering: a feasibility demonstration. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, pages 46–54, New York, NY, USA. ACM.

Zamir, O. and Etzioni, O. (1999). Grouper: a dynamic clustering interface to web search results. In *Proceedings of the eighth international conference on World Wide Web*, WWW '99, pages 1361–1374, New York, NY, USA. Elsevier North-Holland, Inc.

Zeng, H.-J., He, Q.-C., Chen, Z., Ma, W.-Y., and 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, pages 210–217, New York, NY, USA. ACM.

Zimmerman, M. (2000). Weaving the web: the original design and ultimate destiny of the world wide web by its inventor [book review]. *IEEE Transactions on Professional Communication*, 43(2):217 –218.

VITA

Dragos Anastasiu was born in Bucharest, Romania, on February 13, 1979, the son of Mariana and Miron Anastasiu. He came to the United States to pursue a Bachelor Degree in Theology. After completing his work at Moody Bible Institute in Chicago, Illinois, he worked in the Information Technology industry for a number of years. In Summer 2008 he entered Texas State University-San Marcos. In the Spring of 2009, he received a post-graduate Certificate in Computer Science from Texas State University-San Marcos. He continued on at Texas State University-San Marcos, pursuing a Masters Degree in Computer Science.

Permanent Address: 500 Keystone Loop
Kyle, Texas 78640

This thesis was typed by Dragos Anastasiu.