International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

A New Bisecting K-means Algorithm for Inferring User Search Goals Engine

Agale Deepali B.[1], Chaudhari J.L.[2]

[1]Bhivarabai Savant Instituute of Technology and Research, JSPM , Wagholi, Pune, India

[2]Assistant Professor at Bhivarabai Savant Instituute of Technology and Research, JSPM, Wagholi, Pune, India

Abstract: ***Different users may want to search different goals when they submit some ambiguous query, to a search engine. The inference of user search goals can be very useful in improving performance of search engine. To conclude user search goals by analyzing search engine query logs a novel approach is proposed. First thing is that, we propose a framework to find out different user search goals for a query by clustering the proposed feedback sessions. Feedback session is built from user click-through data and can efficiently reflect the information needs of users. Second thing is, we propose a novel approach to generate pseudo-documents by using feedback sessions for clustering. For clustering a new algorithm which is bisecting K-means algorithm is used. At the end, a new criterion "Classified Average Precision (CAP)" is proposed to evaluate the performance of search enging.This criteria gives us value for k-means and bisecting k-means algorithm which shows that bisecting algorithm has better performance than k-means.***

Keywords: User search goals Feedback Sessions, Pseudo-Documents, Restructuring Search Results, Classified Average Precision

1. Introduction

This approach is used to infer user search goals for a query by clustering proposed feedback sessions. Feedback session is series of clicked and unclicked urls. Clustering algorithm is applied on to pseudo-documents which is generated from feedback session. So it forms clusters according to user search goals or queries. Finally evaluation criterion is used to check the performance of the system. It compares the performance of k-means and bisecting k-means algorithm. Bisecting k-means gives the better performance than k-means.

The objectives of the project are as below:

1. **Feedback Session:** The proposed feedback session consists of both clicked and unclicked URLs and ends with the URL that was clicked in a single session at last. It is motivated that before the last click, all the URLs have been scanned and evaluated by users. In this way it shows the listed clicked and unclicked URL's by user.
2. **Optimization method to map:** feedback session is mapped onto pseudo document, which consists of titles and snippets by using optimization method.

Bisecting K-Means Clustering Algorithm

To cluster pseudo document clustering algorithm is applied.

2. Evaluation Criterion

To check performance of system an evaluation criterion is used

3. Related Work

In recent years, many works have been done to infer the socalled user goals or intents of a query [14], [15], [18]. But in fact, their works belong to query classification. Some works analyze the search results returned by the search engine directly to exploit different query aspects [7], [21]. However, query aspects without user feedback have limitations to improve search engine relevance. Some works take user feedback into account and analyze the different clicked URLs of a query in user click-through logs directly, nevertheless the number of different clicked URLs of a query may be not big enough to get ideal results. Wang and Zhai clustered queries and learned aspects of these similar queries [19], which solves the problem in part. However, their method does not work if we try to discover user search goals of one single query in the query cluster rather than a cluster of similar queries. For example, in [18], the query "car" is clustered with some other queries, such as "car rental," "used car," "car crash," and "car audio." Thus, the different aspects of the query "car" are able to be learned through their method. However, the query "used car" in the cluster can also have different aspects, which are difficult to be learned by their method. Some other works introduce search goals and missions to detect session boundary hierarchically [12]. However, their method only identifies whether a pair of queries belong to the same goal or mission and does not care what the goal is in detail. A prior utilization of user click-through logs is to obtain user implicit feedback to enlarge training data when learning ranking functions in information retrieval. Thorsten Joachims did many works on how to use implicit feedback to improve the retrieval quality [9], [10], [11]. They consider feedback sessions as user implicit feedback and propose a novel optimization method to combine both clicked and unclicked URLs in feedback sessions to find out what users really require and what they do not care. One application of user search goals is restructuring web search results. There are also some related works focusing on organizing the search results [7], [8], [21]. They infer user search goals from user click-through logs and restructure the search results according to the inferred user search goals.

Many works about user search goals analysis have been investigated. They can be summarized into three classes:

Licensed Under Creative Commons Attribution CC BY

query classification, search result reorganization, and session boundary detection.

S. BEITZEL, E. JENSEN, A. CHOWDHURY, AND O. FRIEDER, "VARYING APPROACHES TO TOPICAL WEB QUERY CLASSIFICATION," PROC. 30TH ANN. INT'L ACM SIGIR CONF. RESEARCH AND DEVELOPMENT (SIGIR '07), PP. 783-784, 2007.

In this first class, topical classification [5]of web queries has drawn recent interest because of the promise it offers in improving retrieval effectiveness and efficiency. However, much of this promise depends on whether classification is performed before or after the query is used to retrieve documents. We examine two previously unaddressed issues in query classification: pre vs. post-retrieval classification effectiveness and the effect of training explicitly from classified queries vs. bridging a classifier trained using a document taxonomy. Bridging classifiers map the categories of a document taxonomy onto those of a query classification problem to provide sufficient training data. We find that training classifiers explicitly from manually classified queries outperforms the bridged classifier by 48% in F1 score. Also, a pre-retrieval classifier using only the query terms performs merely 11% worse than the bridged classifier which requires snippets from retrieved documents. Thus people attempt to infer user goals and intents by predefining some specific classes and performing query classification accordingly. However, since what users care about varies a lot for different queries, finding suitable predefined search goal classes is very difficult and impractical.

B. POBLETE AND B.-Y RICARDO, "QUERY-SETS: USING IMPLICIT FEEDBACK AND QUERY PATTERNS TO ORGANIZE WEB DOCUMENTS," PROC. 17TH INT'L CONF. WORLD WIDE WEB (WWW '08), PP. 41-50, 2008.

In the second class, Effective organization of search results[19] is critical for improving the utility of any search engine. Clustering search results is an effective way to organize search results, which allows a user to navigate into relevant documents quickly. However, two deficiencies of this approach make it not always work well: (1) the clusters discovered do not necessarily correspond to the interesting aspects of a topic from the user's perspective; and (2) the cluster labels generated are not informative enough to allow a user to identify the right cluster. In this paper, we propose to address these two deficiencies by (1) learning \interesting aspects" of a topic from Web search logs and organizing search results accordingly; and (2) generating more meaningful cluster labels using past query words entered by users. We evaluate our proposed method on a commercial search engine log data. Compared with the traditional methods of clustering search results, our method can give better result organization and more meaningful labels. people try to reorganize search results. But this involves many noisy search results that are not clicked by any users. In the third class, people aim at detecting session boundaries. However, this only identifies whether pair of queries belongs to the same goal and does not care what the goal is in detail.

Table 1: Previous Work

Previous Research Papers	*Result/Conclusion*
Z. Chen	Worked on Query classification Limitations- Experiment was conducted on a potentially-biased dataset
H. Chen	Organizes search results into a hierarchical category structure. Limitations- Query aspects without user feedback have limitations to improve search engine relevance
Wang , Zhai	clustered queries and learned aspects of similar queries Limitations- This method does not work if we try to discover user search goals of any one single query in the query cluster rather than a cluster of similar queries.
R. Jones and K.L. Klinkner,	Introduce search goals and missions to detect session boundary hierarchically Limitations- Their method only identifies whether a pair of queries belong to the same goal or not and does not care what the goal is in detail.

4. Proposed Work

The overall system architecture is as shown in figure 3.1 below. First block of figure displays the feedback session of search results. Then by using optimization method titles and snippets are extracted from feedback session, this collection of titles and snippets is called as pseudo document. After that clustering algorithm applied on to these pseudo document. Finally clustered data is obtained and at the same time by using evaluation criteria performance of algorithm is checked.

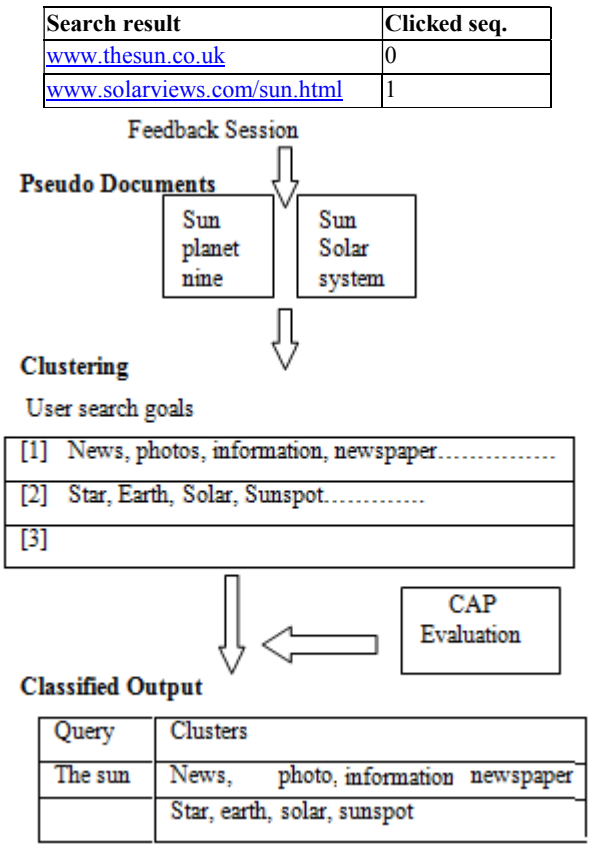


Figure 3.1: Framework of our approach

Licensed Under Creative Commons Attribution CC BY

4.1 Conversion Of Feedback Sessions To Pseudo-Documents

Building of pseudo-document has two steps.

4.2.1 Representation of URLs In the Feedback Session.
In the first step, we first enrich the URLs with additional textual contents by extracting the titles and snippets of the returned URLs appearing in the feedback session. In such way, each URL in the feedback session is represented by a small text paragraph that consists of that URLs title and snippet. After that some textual processes are implemented to those text paragraphs, such as transforming all of the letters to lowercases, stemming and removing stop words. Finally, each URL's title and snippet are represented by a Term Frequency-Inverse
Document Frequency (TF-IDF) vector, respectively, as in

$Ti = [tw_1, tw_2, tw_3 \ldots\ldots twn]^T$
$Si = [Sw_1, Sw_2, Sw_3 \ldots\ldots Swn]^T$
(1)
Where Ti and Sj are the TF-IDF vectors of the URL's title and snippet, respectively. wj= (1,2,3,……n) is the jth term appearing in the enriched URLs. Here, a "term" is defined as a word or a number in the dictionary of document collections. twj and swj represent the TF-IDF value of the jth term in the URL's title and snippet, respectively. Considering that each URLs' titles and snippets have different significances, we represent the each enriched URL by the weighted sum of Tui and Sui, namely

$$Ri = wt \cdot \mathbf{T}i + st \cdot \mathbf{S}i = [f\ w_1, f\ w_2, \ldots\ldots\ldots\ldots, f wn]^T \quad (2)$$

Where Ri means the feature representation of the i^{th} URL in the feedback session, and *w*t and *s*t are the weights of the titles and the snippets, respectively.

4.2.2 Formation of Pseudo-Document
We propose an optimization method to combine clicked and unclicked URLs in the feedback session to obtain a feature representation.
Let R be the feature representation of a feedback session, and (*w*) be the value for the term w.
Let
C = (m=1, 2, 3… M), and
UC = (l=1, 2, 3… L);
Let R be the feature representations of the clicked and unclicked URLs in this feedback session, respectively.
Let C and UC be the values for the term w in the vectors. We want to obtain such a S that the sum of the distances between S and each C is minimized and the sum of the distances between S and each UC is maximized. Based on the assumption that the terms in the vectors are independent, we can perform optimization on each dimension independently, as shown in below equation.

$$\mathbf{S} = [ff(w\square), ff(w\square), ff(w\square), \ldots\ldots\ldots, ff(wn)]^T \quad (3)$$

$$Rs(w) = \arg\min \sum_M \left(S(w) - C(w)\right)^2 - \lambda \sum_L \left((S(w) - UC(w))^2\right) \quad (4)$$

λ is a parameter balancing the importance of clicked and unclicked URLs. When λ in (4) is 0, unclicked URLs are not taken into account. On the other hand, if λis too big, unclicked URLs will dominate the value of Uc. In this project, we set λ to be 0.5.

4.3 Clustering of Pseudo Document

As in equation (3) and (4), each feedback session is represented by a pseudo-document and the feature representation of the pseudo-document is R*s*. The similarity between two pseudo-documents is computed as the cosine score of Rs_i and R*s*j, as follows:

$$Simij = \cos(Rs_i, Rsj) = \frac{\mathbf{Rs_i} \cdot \mathbf{Rsj}}{|\mathbf{Rs_i}| \cdot |\mathbf{Rsj}|} \quad (5)$$

And distance between two feedback sessions is calculated by using formula

$$Disij = 1 - Simij$$

To cluster pseudo documents K-means clustering is used which is very simple and effective. To check the optimal values of clustering we have a evaluation criterion.

4.4 Bisecting Algorithm

For Bisecting algorithm you must cluster documents using k-means algorithm and then on the result of k-means algorithm you can apply bisecting algorithm.

Read following bisecting steps.
The idea is iteratively splitting your cloud of points in 2 parts. In other words, you build a random binary tree where each splitting (a node with two children) corresponds to splitting the points of your cloud in You begin with a cloud of points.

- Compute its centroid (barycenter) w
- Select randomly a point cL among the points of the cloud
- Construct point cR as the symmetric point of cL when compared to w (the segment cL->w is the same as w->cR)
- Separate the points of your cloud in two, the ones closest to cR belong to the subcloud R, and the ones closest to cL belongs to the subcloud L
- Reiterate for the subclouds R and L

Notes :
You can discard the random points once you've used them already. However, keep the centroids of all the subcoulds.
Stop at point when your subclouds contain exactly one point.

5. Design Process

5.1 Data Flow Diagram

A data flow diagram (DFD) is a graphical representation of the "flow" of data through an information system. It differs

Licensed Under Creative Commons Attribution CC BY

from the flowchart because it shows the data flow instead of the control flow of the program. It can also be used for the visualization of data processing. The data flow diagram is designed to show how a system is divided into smaller portions and to highlight the flow of data between those parts.

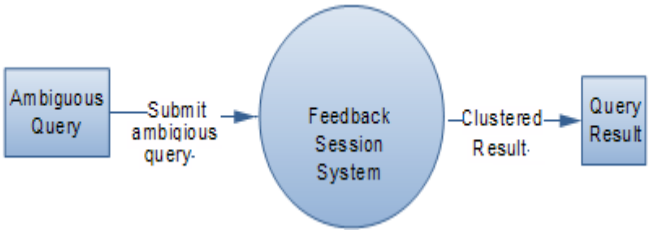


Figure 5.1:.1DFD 1

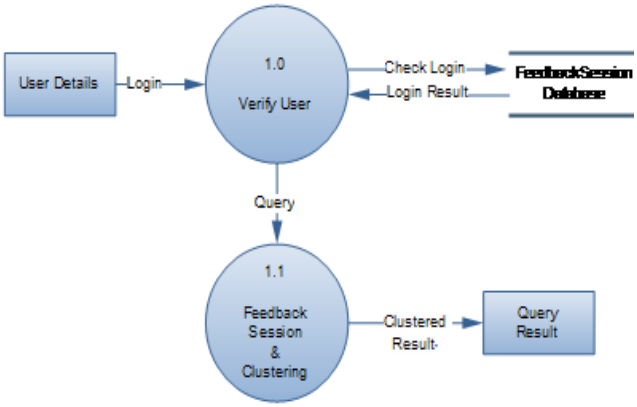


Figure 5.2: DFD 2

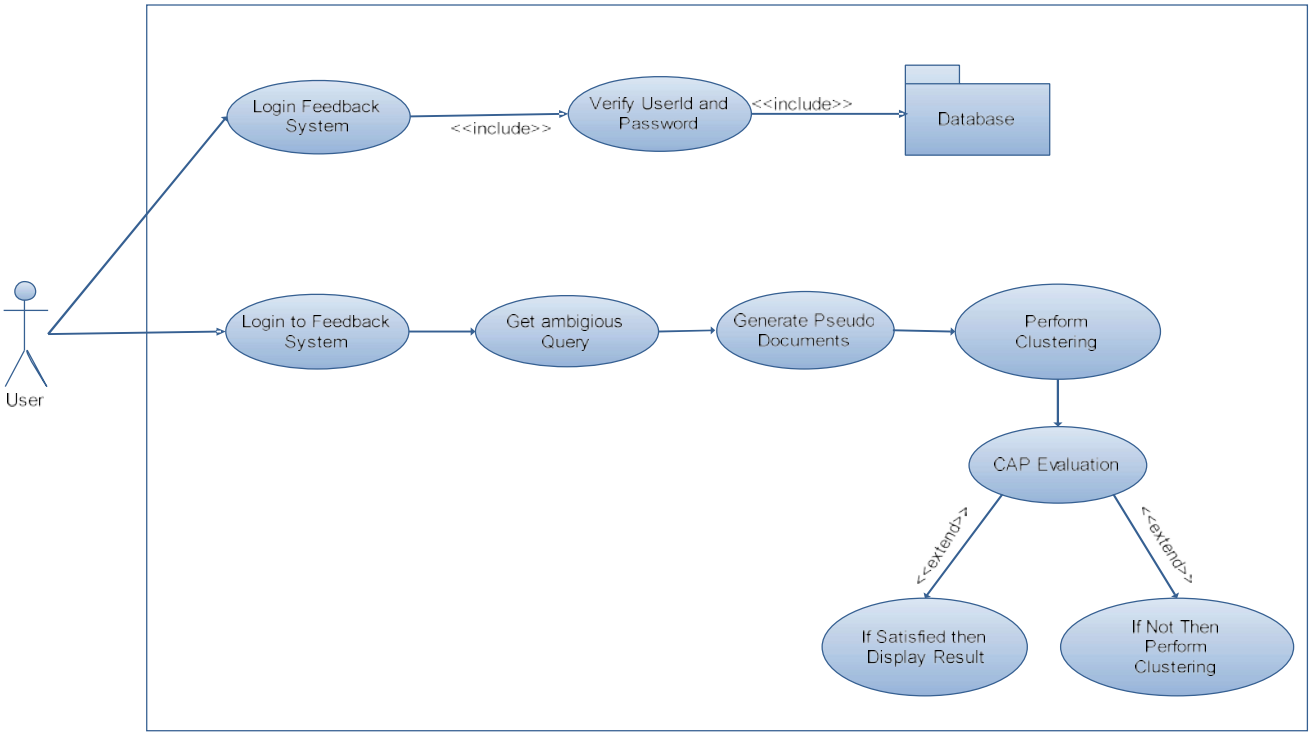


Figure 5.5.1: Use case diagram

5.2 UML Diagrams

The Unified Modeling Language (UML) is a graphical language for visualization, specifying, constructing and documenting the artifacts of a software intensive system. The UML gives a standard was to write systems blue prints, covering conceptual things, such as business processes and system functions, as well as concrete things, such as classes written in a specific programming language, database schemas, and reusable software components.

5.2.1 Use Case Diagram

It shows a set of use cases and actors (a special kind of class and their relationships). Use case diagrams address the static use case view of system. These diagrams are especially important in organizing and modeling the behavior of the system.

6. Evaluation Criterion

6.1 Average Precision

A possible evaluation criterion is the average precision (AP) which evaluates according to user implicit feedbacks. AP is the average of precisions which is computed at the point of each relevant document in the ranked sequence, shown in

$$AP = \frac{1}{N^+} \sum_{r=1}^{N} rel(r) \frac{R_r}{r} \quad (6)$$

Where

N is the number of relevant (or clicked) documents in the retrieved ones,

Licensed Under Creative Commons Attribution CC BY

r is the rank, N is the total number of retrieved documents, rel(r) is a binary function on the relevance of a given rank, and
Rr is the number of relevant retrieved documents of rank r or less.

6.2 Voted AP (VAP)

It is calculated for purpose of restructuring of search results classes i.e. different clustered results classes. It is same as AP and calculated for class which having more clicks.

6.3 Risk

It is the AP of the class including more clicks? There should be a risk to avoid classifying search results into too many classes by error. So we propose the **Risk**.

$$\text{Risk} = \frac{\sum_{i,i=1(i<j)}^{m} dij}{C^2 m}$$

6.4 Classified AP (CAP)

VAP is extended to CAP by introducing combination of VAP and Risk. Classified AP can be calculated by using the formula, as follows:

$$\text{CAP} = \text{VAP} \times (1 - Risk)^{\gamma}$$

7. Original Results

The goal of paper is to infer user search goals by maintaining feedback session. Then from feedback session pseudo document is generated on to which clustering algorithm is applied. Initially k-means clustering algorithm is applied. Evaluation criterion is defined to check the performance of the system. Then bisecting k means algorithm is applied on to same pseudo document. Finally performance of these two algorithms is checked by using evaluation criterion. Bisecting k-means algorithm gives higher values than k-means clustering. The comparison of two algorithm is as shown with the help of graphs. When user enters search query, first of all system will generates the id for each query. Figure 9.1 shows queries with their related id's.

QId	Query
316	moon
318	circle
319	coffee
320	cricket
321	company
322	map
326	diamonds
327	earth
328	lamborghini
329	mickle jackson
330	mission
331	networks
332	school
333	apple
334	presentation
335	search
337	fruits
350	king1
354	king

Figure 9.1 Queries and related id's

Then system will display the search results for query. For each query system will maintained the clicked and unclicked sequence of url's, it means that it maintains the feedback session for each query. Figure 9.2 shows Google search result with feedback session.

GId	QId	Title	Snippet	URL	IsClicked	Seq	ClickSeq	Cluster
5196	318	1) Circle Interne...	Circle is a Bitcoin...	https://www.cir...	False	1	0	Cluster1
5197	318	2) Circle - Wikipe...	A circle is a simpl...	http://en.wikipe...	True	2	1	Cluster2
5198	318	3) Circle – Conc...	Find best things ...	http://discoverci...	False	3	0	Cluster2
5199	318	4) Circle - Math i...	circle. A circle is ...	http://www.mat...	True	4	2	Cluster2
5200	318	5) CIRCLE	In 2013, CIRCLE...	http://www.civic...	False	5	0	Cluster1
5201	318	6) Circle -- from ...	the area of the ...	http://mathworl...	True	6	3	Cluster2
5202	318	1) Circle Interne...	Circle is a Bitcoin...	https://www.cir...	True	1	1	NULL
5203	318	2) Circle - Wikipe...	A circle is a simpl...	http://en.wikipe...	False	2	0	NULL
5204	318	3) Circle – Conc...	Find best things ...	http://discoverci...	True	3	2	NULL
5205	318	4) Circle - Math i...	circle. A circle is ...	http://www.mat...	False	4	0	NULL
5206	318	5) CIRCLE	In 2013, CIRCLE...	http://www.civic...	True	5	3	NULL
5207	318	6) Circle -- from ...	the area of the ...	http://mathworl...	False	6	0	NULL
5208	318	7) Find & add pe...	Share the right t...	https://support....	True	7	4	NULL
5209	319	1) Coffee - Wiki...	Coffee is a brew...	http://en.wikipe...	True	1	1	Cluster2
5210	319	2) Green Mount...	K-Cup and Vue p...	http://www.gre...	False	2	0	Cluster1
5211	319	4) Stumptown C...	Locally owned c...	http://stumptow...	False	4	0	Cluster2
5212	319	5) Coffee | Star...	Since 1971, Star...	http://www.star...	True	5	3	Cluster2
5213	319	7) Caribou Coffee	Coffee Houses l...	https://www.car...	True	7	4	Cluster2
5214	320	1) Cricket - Wiki...	Cricket is a bat-...	http://en.wikipe...	True	1	1	Cluster1

Figure 9.2: Google search Results with Feedback Session

Licensed Under Creative Commons Attribution CC BY

From feedback session pseudo document is generated .Pseudo documents consists of snippets and titles. Then clustering algorithm will be applied on to this pseudo document. It forms the list of keywords for each cluster according to relevance of keywords. Figure 9.3 shows list of keywords for particular cluster. For example ,for query id 318 it shows 10 keywords . Clustering algorithms will be applied on to these keywords.

The performance of the algorithm is checked by calculating average precision for both algorithms. Classified Average Precision (CAP), Voted Average Precision (VAP) ,Risk are calculated for k-means algorithm and bisecting k-means algorithm. Figure 9.4 shows the values of CAP, VAP and Risk.

	id	qid	VAP	CA
▶	33	318	1	1
	34	318	1	1
	35	319	0.33333333333...	0.3
	36	319	0.55555555555...	0.4182766476...
	39	321	0.16598639455...	0.10217661117...
	40	321	1	0.6
	41	322	0.55555555555...	0.5
	42	322	1	1
	43	323	0.40432098765...	0.4
	44	323	0.55	0.3
	47	325	1	0.6
	48	325	1	0.6
	49	328	0.20833333333...	0.2
	50	328	0.25	0.2
	51	329	0.28	0.2
	52	329	0	0
	53	330	0.47916666666...	0.2
	54	330	0	0

Figure 9.4: CAP, VAP, Risk for k-means and bisecting k-means algorithm

The performance of the system is shown by graphs. Figure 9.5 shows the graph for CAP against query id. And figure 9.6 shows graph for risk against VAP.

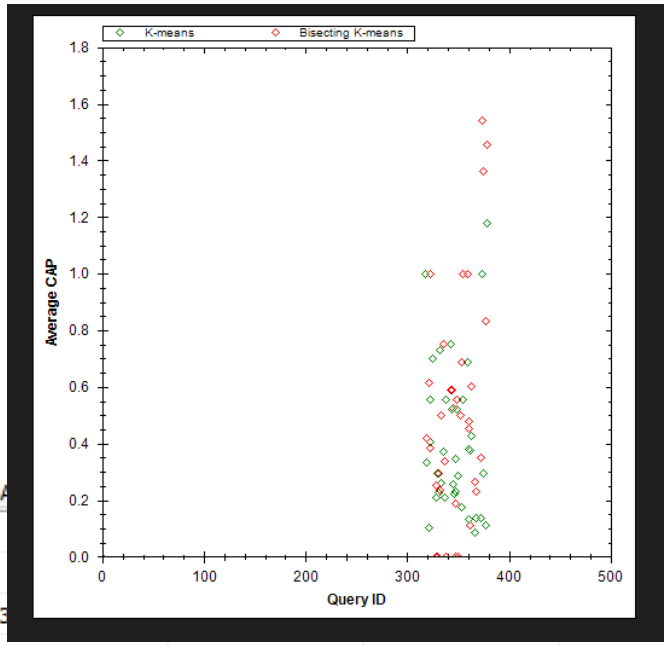


Figure 9.5: Graph for Query ID against CAP

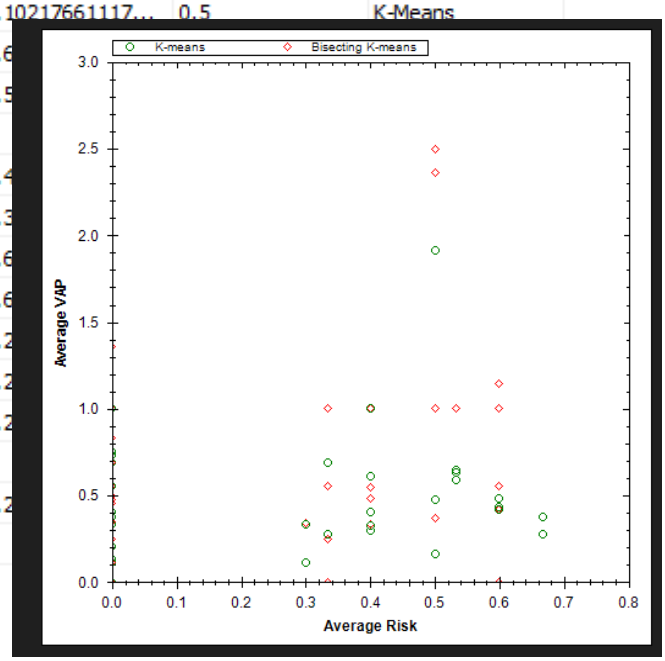


Figure 9.6: Graph for Risk and VAP

8. Conclusion

The proposed system can be used to improve discovery of user search goals for a similar query by using bisecting algorithm for clustering user feedback sessions represented by pseudo-documents. By using proposed system, the inferred user search goals can be used to restructure web search results. So, users can find exact information quickly and very efficiently. The discovered clusters of query can also be used to assist users in web search.

9. Acknowledgment

Ms. Deepali Agale is thankful to Prof. J.L. Chaudhari, Asst. Professor, Information Technology Department, Bhivarabai Sawant Institute of Technology and Research, Wagholi, Pune, for her constant support and helping out with the

Licensed Under Creative Commons Attribution CC BY

preparation of this paper. Also thankful to the Principal, Dr. D. M. Yadav Bhivarabai Sawant Institute Of Technology and Research ,Wagholi,Pune, and Prof. G.M.Bhandari, HOD, Computer Engineering Department Bhivarabai Sawant Institute Of Technology and Research ,Wagholi, Pune, for being a constant source of inspiration.

10. Other Recommendations

Equalize the length of your columns on the last page. If you are using *Word*, proceed as follows: Insert/Break/Continuous.

References

[1] Zheng Lu, Student Member, IEEE, Hongyuan Zha, Xiaokang Yang, Senior Member, IEEE, Weiyao Lin, Member, IEEE, and Zhaohui Zheng**,"** A New Algorithm for Inferring User Search Goals with Feedback Sessions"

[2] S. Beitzel, E. Jensen, A. Chowdhury, and O. Frieder, "Varying Approaches to Topical Web Query Classification," Proc. 30th Ann. Int'l ACM SIGIR Conf. Research and Development (SIGIR '07), pp. 783-784, 2007.

[3] B. Poblete and B.-Y Ricardo, "Query-Sets: Using Implicit Feedback and Query Patterns to Organize Web Documents," Proc. 17th Int'l Conf. World Wide Web (WWW '08), pp. 41-50, 2008.

[4] D. Shen, J. Sun, Q. Yang, and Z. Chen, "Building Bridges for Web Query Classification," Proc. 29th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '06), pp. 131-138, 2006.

[5] H. Chen and S. Dumais, "Bringing Order to the Web: Automatically Categorizing Search Results," Proc. SIGCHI Conf. Human Factors in Computing Systems (SIGCHI '00), pp. 145-152, 2000.

[6] X. Wang and C.-X Zhai, "Learn from Web Search Logs to Organize Search Results," Proc. 30th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '07), pp. 87-94, 2007.

[7] R. Jones and K.L. Klinkner, "Beyond the Session Timeout: Automatic Hierarchical Segmentation of Search Topics in Query Logs," Proc. 17th ACM Conf. Information and Knowledge Management (CIKM '08), pp. 699-708, 2008.

[8] R. Baeza-Yates, C. Hurtado, and M. Mendoza, "Query Recommendation Using Query Logs in Search Engines," Proc. Int'l Conf. Current Trends in Database Technology (EDBT '04), pp. 588-596, 2004.

[9] D. Beeferman and A. Berger, "Agglomerative Clustering of a Search Engine Query Log," Proc. Sixth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD '00), pp. 407-416, 2000.

[10] M. Pasca and B.-V Durme, "What You Seek Is what You Get: Extraction of Class Attributes from Query Logs," Proc. 20th Int'l Joint Conf. Artificial Intelligence (IJCAI '07), pp. 2832-2837, 2007.

[11] A. Bonnaccorsi, "On the Relationship between Firm Size and Export Intensity," Journal of International Business Studies, XXIII (4), pp. 605-635, 1992. (journal style)

[12] R. Caves, Multinational Enterprise and Economic Analysis, Cambridge University Press, Cambridge, 1982. (book style)

[13] M. Clerc, "The Swarm and the Queen: Towards a Deterministic and Adaptive Particle Swarm Optimization," In Proceedings of the IEEE Congress on Evolutionary Computation (CEC), pp. 1951-1957, 1999. (conference style)

[14] H.H. Crokell, "Specialization and International Competitiveness," in Managing the Multinational Subsidiary, H. Etemad and L. S, Sulude (eds.), Croom-Helm, London, 1986. (book chapter style)

[15] K. Deb, S. Agrawal, A. Pratab, T. Meyarivan, "A Fast Elitist Non-dominated Sorting Genetic Algorithms for Multiobjective Optimization: NSGA II," KanGAL report 200001, Indian Institute of Technology, Kanpur, India, 2000. (technical report style)

[16] J. Geralds, "Sega Ends Production of Dreamcast," vnunet.com, para. 2, Jan. 31, 2001. [Online]. Available: http://nl1.vnunet.com/news/1116995. [Accessed: Sept. 12, 2004]. (General Internet site)

Author Profile



Agale Deepali Bansi received the B.E. from Padmashri Dr. Vikhe Patil College of Engg. From IT department and pursuing my M.E. degree from bhivarabai Savant Institute of Technology and Research, JSPM, Wagholi, Pune in computer Engineering. And now I am working as assistant professor in Shri Chatrapati Shivaji Maharaj College of Engineering, Nepti, Ahmednagar.

Licensed Under Creative Commons Attribution CC BY