International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

Fast Prototyping Networks and Using Datamining Applications

Tewhasom Aregay Weldu

Lecturer/HOD, Department of Computer Science, Adigrat University, Ethiopia

Abstract: ***To study different applications of mining for the fast prototyping networks and improvement of search engine ranking, network information retrieval and fast prototyping network environment enhancement. The advantage of the implicit feedback left in the trail of users while navigating through fast prototyping networks. The value of queries to extract interesting rules, patterns and information about the documents they reach. The models, created in this Paper work, show that the "wisdom of the crowds" conveyed in queries has many applications that overall provide a better understanding of user's needs in the fast prototyping network. The general interaction of applications with fast prototyping networks and searching engines in a straightforward way. We focus our efforts on analyzing and extracting valuable knowledge from the behavior of users on the fast prototyping networks. Much of this information is provided implicitly by users and recorded in usage logs, which include search engine query logs and/or network access logs. In particular, we center our research on queries that users submit to fast prototyping search engines, which we believe convey in straightforward way "wisdom of the crowds". The intuition is that queries and their clicked results implicitly convey the opinion of users about specific network documents. As discuss throughout this paper, queries are crucial to understanding how users interact with Web sites and search engines. Implicit user feedback provides a unique insight into users' actual needs on the Web.***

Keywords: Wisdom of the crowds, fast prototyping, search engines, websites, Network data mining

1. Introduction

The fast prototyping network is unlike any other repository of information that we have ever studied before; it is an immensely rich repository which grows at an astoundingly fast pace. These unique characteristics carry many new challenges for Fast prototyping network researchers, which include among other things, high data dimensionality and highly volatile and constantly evolving content. Due to this, it has become increasingly necessary to create new and improved approaches to traditional data mining techniques can be applied to the Fast prototyping network.

In this regard, recognizing and separating automatically interesting and valuable information, has become a very relevant problem when processing such huge quantities of data. The key issues in this matter are: how do we know which information is interesting or useful? and how can we find this information automatically?.In this Paper we focus our efforts on analyzing and extracting valuable knowledge from the behavior of users on the Fast prototyping network.

The work on queries that users submit to fast prototyping network search engines, which we believe convey in straightforward way "wisdom of the crowds". The intuition is that queries and their clicked results implicitly convey the opinion of users about specific Fast prototyping network documents.

2. Background

Fast prototyping network mining, specifically on the topics of Fast prototyping network usage mining and query mining. To do this, first we will discuss some preliminary data mining techniques, such as clustering and frequent itemset mining, as well as some relevant concepts, such as the vector space model.

Clustering

Cluster analysis is a technique used to group data into sets of elements that are meaningful and/or useful. There are several types of clustering techniques available. In this work we use bisecting k-means. This is a straightforward extension of the k-means algorithm. This extension consists of a k-way clustering solution generated by a sequence of k–1repeated bisections. There are several global clustering criterion functions that can be used to select which cluster to bisect next, in the clustering process. In this work we use I1, I2, H1 and H2, as defined in [8]. The formulas for these functions are:

$$\mathcal{I}_1 = \sum_{i=1}^{k} \frac{1}{n_i} \left(\sum_{v,u \in S_i} sim(v,u) \right),$$

$$\mathcal{I}_2 = \sum_{i=1}^{k} \frac{1}{n_i} \sqrt{\sum_{v,u \in S_i} sim(v,u)},$$

$$\mathcal{H}_1 = \mathcal{I}_1 \sum_{i=1}^{k} \frac{1}{n_i} \frac{\sqrt{\sum_{v,u \in S_i} sim(v,u)}}{\sum_{v \in S_i, u \in S} sim(v,u)},$$

$$\mathcal{H}_2 = \mathcal{I}_2 \sum_{i=1}^{k} \frac{1}{n_i} \frac{\sqrt{\sum_{v,u \in S_i} sim(v,u)}}{\sum_{v \in S_i, u \in S} sim(v,u)}$$

Frequent Itemset Mining

An itemset is a collection of zero or more items that occur together within a same transaction. More formally, let I = {i1, i2... id} be the set of all possible items in a data collection, and let T = {t1, t2, . . . ,tN} be the set of all transactions. Each transaction tjcontains a subset of items, or itemset, from I.

Fast prototyping network Usage Mining for Fast prototyping network Terminal Improvement

Fast prototyping network usage mining has generated a great amount of commercial interest [4]. There is an extensive list of work that incorporates Fast prototyping network usage mining for several purposes. One of its applications is in the improvement of Fast prototyping network terminals, which has been mostly focused on the support of "adaptive Fast prototyping network terminals" and "automatic personalization" [5]. The main purpose of these types of applications is to find interesting rules and patterns in the usage data of Fast prototyping network terminals by using data mining techniques such as: analysis of frequent navigational patterns, document clustering, and association rules, based on the pages terminal by users [2].

Query-Sets: Using Implicit Feedback and Query Patterns to Organize Fast prototyping network Documents

As the amount of contents in the Fast prototyping network grow, it becomes increasingly difficult to manage and classify its information. Optimal organization of Fast prototyping network documents is important for Fast prototyping network terminals as well as for heterogeneous sets of documents.

Document Clustering and Labeling

There are two main data sources for obtaining clicked queries for documents, and depending on the source we might have partial queries or complete queries:

Partial queries: when organizing general Fast prototyping network documents or search results. Query clicks to documents discovered from this log are only the ones that were submitted to the particular search engine that generated the log. Therefore, the more widely used the search engine is, the better it will represent the real usage of documents.

Complete queries: In Fast prototyping network terminals access logs. This situation is most likely when organizing documents belonging to a particular Fast prototyping network terminal. Standard combined access logs allow (very easily) discovering all of the queries from Fast prototyping network search engines that directed traffic to the terminal (i.e., queries from which documents in the terminal were clicked). This log may also contain information about queries to the internal search engine of the Fast prototyping network terminal (if one is available).

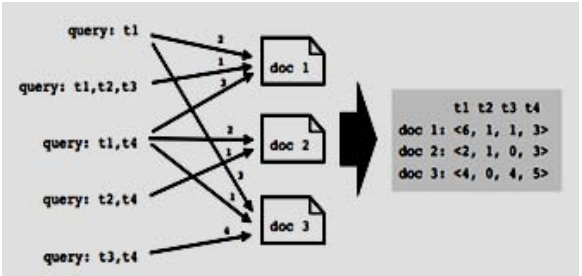

Figure 1: Query document representation, without normalization

Discovering Similar Fast prototyping network Terminals Using Search Engine Queries

Fast prototyping network IR from the traditional retrieval of Fast prototyping network documents that satisfy a certain query, towards the retrieval of complete Fast prototyping network terminals.

"Fast prototyping network terminal" and then go ahead to model Fast prototyping network terminals as vectors, by extending the traditional vector space model for documents. Our framework is generic and allows different Fast prototyping network terminal models, but our focus is on modeling a terminal as a vector over a query-based feature space. These methods to build and reduce the size of query-based feature spaces.

Clustering Fast prototyping network Terminals

Fast prototyping network terminals, we apply existing clustering techniques, using the vector representation generated by each model to generate different solutions, for the global clustering functions I1 , I2, H1 and H2

External Cluster Quality Measures

We present the results obtained for the different clustering solutions regarding an external cluster quality indicator, in this case DMOZ categories. In this study, we consider the DMOZ categories to be the real categories of the Fast prototyping network terminals. Therefore, we measure the quality of the clustering solutions against this "gold standard".

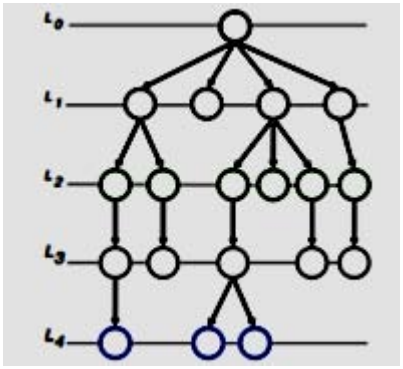

Figure 2: Hierarchy level of directory

The quality of each clustering solution is measured using the solution's entropy and purity.

Table 1: Purity values for I1, I2, H1 and H2

Purity(I_1)					
No. of Clusters	*46*	*75*	*104*	*216*	*Avg.*
FULLPATTERNS	0.732	0.711	0.600	0.514	0.639
FULLQUERIESPLUS	0.757	0.718	0.591	0.545	**0.653**
QUERYTERMS	0.744	0.710	0.596	0.526	0.644
TEXT	0.720	0.699	0.582	0.571	0.643
Purity(I_2)					
No. of Clusters	*46*	*75*	*104*	*216*	*Avg.*
FULLPATTERNS	0.734	0.697	0.566	0.516	0.628
FULLQUERIESPLUS	0.749	0.716	0.594	0.550	**0.652**
QUERYTERMS	0.739	0.686	0.581	0.528	0.634
TEXT	0.724	0.700	0.580	0.548	0.638

Purity(H_1)					
No. of Clusters	46	75	104	216	Avg.
FULLPATTERNS	0.703	0.683	0.577	0.508	0.618
FULLQUERIESPLUS	0.731	0.722	0.588	0.554	**0.649**
QUERYTERMS	0.733	0.671	0.576	0.519	0.625
TEXT	0.724	0.696	0.601	0.554	0.644
Purity(H_2)					
No. of Clusters	46	75	104	216	Avg.
FULLPATTERNS	0.718	0.674	0.550	0.498	0.610
FULLQUERIESPLUS	0.723	0.691	0.571	0.527	0.628
QUERYTERMS	0.716	0.679	0.561	0.514	0.618
TEXT	0.728	0.701	0.580	0.532	**0.635**

Table 2: Entropy Values of I1, I2, H1 and H2

Purity(I_2)					
No. of Clusters	46	75	104	216	Avg.
FULLQUERIESPLUS	0.794	0.773	0.718	0.773	**0.765**
FULLQUERIES	0.813	0.770	0.723	0.746	0.763
MAXPATTERNS	0.784	0.763	0.694	0.739	0.745
TEXT	0.790	0.803	0.691	0.731	0.754
Entropy(I_2)					
No. of Clusters	46	75	104	216	Avg.
FULLQUERIESPLUS	0.138	0.124	0.132	0.067	**0.115**
FULLQUERIES	0.159	0.137	0.133	0.072	0.125
MAXPATTERNS	0.189	0.148	0.149	0.075	0.140
TEXT	0.140	0.107	0.137	0.074	**0.115**

Table 3: Purity and Entropy values for the reduced set of web sites using I2

Entropy(I_1)					
No. of Clusters	46	75	104	216	Avg.
FULLPATTERNS	0.269	0.235	0.277	0.210	0.248
FULLQUERIESPLUS	0.205	0.183	0.238	0.186	**0.203**
QUERYTERMS	0.256	0.222	0.256	0.188	0.231
TEXT	0.257	0.211	0.239	0.167	0.219
Entropy(I_2)					
No. of Clusters	46	75	104	216	Avg.
FULLPATTERNS	0.214	0.194	0.234	0.189	0.208
FULLQUERIESPLUS	0.187	0.176	0.220	0.182	**0.191**
QUERYTERMS	0.214	0.195	0.226	0.181	0.204
TEXT	0.206	0.182	0.214	0.167	**0.192**
Entropy(H_1)					
No. of Clusters	46	75	104	216	Avg.
FULLPATTERNS	0.234	0.205	0.232	0.189	0.215
FULLQUERIESPLUS	0.192	0.174	0.221	0.180	**0.192**
QUERYTERMS	0.215	0.202	0.230	0.184	0.208
TEXT	0.206	0.185	0.210	0.164	**0.191**
Entropy(H_2)					
No. of Clusters	46	75	104	216	Avg.
FULLPATTERNS	0.223	0.199	0.238	0.188	0.212
FULLQUERIESPLUS	0.201	0.188	0.226	0.187	0.201
QUERYTERMS	0.221	0.199	0.228	0.183	0.208
TEXT	0.202	0.177	0.217	0.172	**0.192**

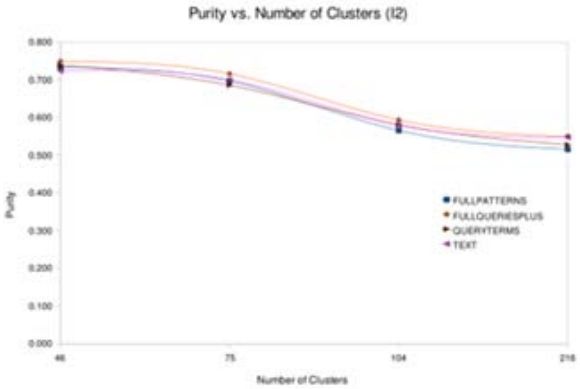

Figure 3: Purity Vs the Number of Clusters for I2

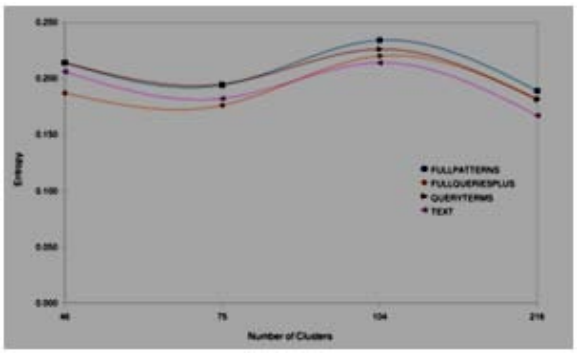

Figure 4: Entropy Vs Number of Clusters for I2

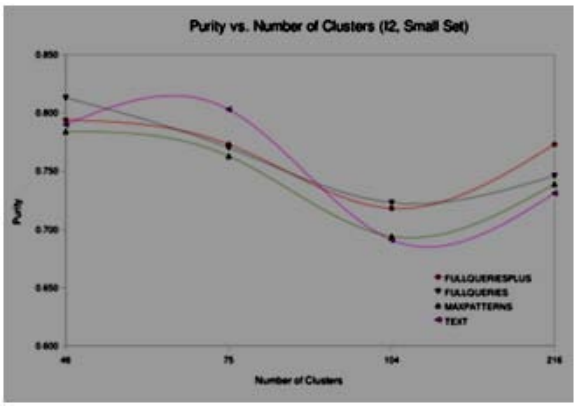

Figure 5: Purity Vs the Number of Clusters for I2

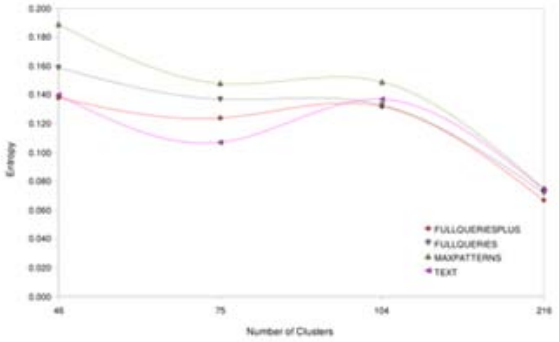

Figure 6: Entropy Vs Number of Clusters for I2

A Fast prototyping network Terminal Mining Model Centered on User Queries

Fast prototyping network terminal search, and not with the intention of discovering new data to increase the quality of the Fast prototyping network terminal. Our generates a visualization of the terminal's content distribution in relation to the link organization between documents, as well as the IP OR IDENTITYs selected due to queries. Fast prototyping network terminals designed that register traffic from internal and/or external search engines, even if this is not the main mechanism of navigation in the terminal. The output of the model consists of several reports from which improvements can be made to the Fast prototyping network terminal.

In our model the structure of the Fast prototyping network terminal is obtained from the links between documents and the content is the text extracted from each document.

External queries: These are queries submitted on Fast prototyping network search engines, from which users selected and viterminald documents in a particular Fast

prototyping network terminal. They can be discovered from the log's referrer field.

Internal queries: These are queries submitted to a Fast prototyping network terminal's internal search box. Additionally, external queries that are specified by users for a particular terminal will be considered as internal queries for that terminal.

3. Query Classification

We classify queries in relation to: if the user chooses to visit the generated results and if the query had results in the Fast prototyping network terminal. Our classification can be divided into two main groups: successful queries and unsuccessful queries. Successful queries can be found both in internal and external queries, but unsuccessful queries can only be found for internal queries since all external queries in the Fast prototyping network terminal's usage logs were successful for that terminal.

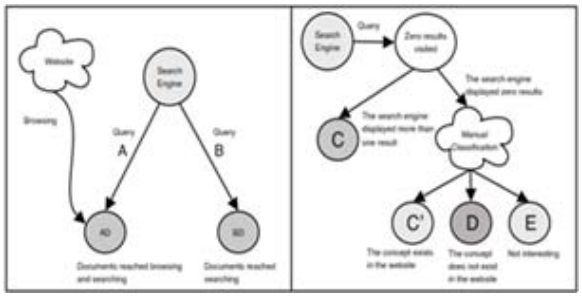

Figure 7: Successful Queries (Right),Unsuccessful queries (Left)

Successful Queries:
There are two types of successful queries, call as A and B. Class A queries: Queries for which the session viterminald one or more results in AD, where AD contains documents found in the DWS set. In other words, the documents in AD have also been reached, in at least one other session, browsing without using a search engine. Class B queries: Queries for which the session viterminald one or more results in BD, where BD contains documents that are only classified as DQ and not in DWS. In other words documents in BD have only been reached using a search in all of the analyzed session.

Unsuccessful Queries:-
There are four types of unsuccessful queries, which we will call C, C', D and E

Class C queries represent concepts that should be developed in depth in the contents of the Fast prototyping network terminal with the meaning that users intended, focused on the keywords of the query.

Class C' queries represent words that should be used in the text that describes links and documents that share the same meaning as these queries.

Class D queries represent concepts that should be included in documents in the Fast prototyping network terminal, because they represent new topics that are of interest to users of the Fast prototyping network terminal.

Class E queries: Queries that are not interesting for the Fast prototyping network terminal, as there are no results, but it's not a class C' or class D query, and should be omitted in the classification

Table 4: Classes of queries and their contribution to the improvement of a web site

Class	Concept exists	Results displayed	Visited documents	Significance	Contribution	Affected component
A	yes	yes	DQ ∩ DWS	low	additional IS	anchor text
B	yes	yes	DQ \ DWS	high	new IS, add hotlinks	anchor text, links
C	yes	yes	∅	medium	new content	documents
C'	yes	no	—	medium	new IS	anchor text, documents
D	no, but it should	no	—	high	new content	anchor text, documents
E	no	no	—	none	—	—

4. Conclusion

In this paper we present the query-set model reduces by over 90% the number of features needed to represent a set of documents and improves by more than 90% the quality. Also, the query-set model shows a higher level of inter-judge agreement which corresponds with the fact. Fast prototyping network terminal models was measured applying clustering to the Fast prototyping network terminal vectors, with the objective of discovering groups of similar terminals. Our experimental evaluation shows that the query-based approaches use significantly less features than the full text approach obtaining better results and also the Fast prototyping network terminal mining model that is focused on query classification. The aim of this model is to find better IS, contents and link structure for a Fast prototyping network terminal. Our tool discovers, in a very simple and straight forward way. Our model can be applied to almost any type of Fast prototyping network terminal, without significant previous requirements, and it can still generate suggestions if there is no internal search engine in the Fast prototyping network terminal.

References

[1] AOL research fast protyping networkterminal, no longer online. http://research.aol.com.
[2] Robert Cooley, Pang-Ning Tan, and Jaideep Srivastava. Fast protyping networksift: the fast protyping network terminal information filter system. In KDD Workshop on Fast protyping network Mining, San Diego, CA. Springer-Verlag, in press, 1999.
[3] Fast protyping network characterization activity. http://www.w3.org/WCA/.
[4] Ricardo Baeza-Yates. Mining the fast protyping network (inspanish). El profesional de la infor-macio´ n (The Information Professional), 13(1):4–10, Jan-Feb 2004.
[5] The fast protyping networkalizer. http://www.fast protyping networkalizer.org.
[6] Ricardo A. Baeza-Yates, Carlos A. Hurtado, and Marcelo Mendoza. Improving search engines by query clustering. JASIST, 58(12):1793–1804, October 2007.
[7] Fast protyping networktrends log analyzer. http://www.fast protyping networktrends.com.
[8] Ying Zhao and George Karypis. Criterion functions for document clustering: Ex- periments and analysis.

Technical report, University of Minnesota, Department of Computer Science / Army HPC Research Center, Minneapolis, MN 55455, 2001.

[9] Bamshad Mobasher, Robert Cooley, and Jaideep Srivastava. Automatic personal- ization based on fast protyping network usage mining. Commun. ACM, 43(8):142–151, 2000.