IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 3, No 9, May 2010
ISSN (Online): 1694-0784
ISSN (Print): 1694-0814

A Theoretical Approach to Link Mining for personalization

K.Srinivas[1], L.Kiran Kumar Reddy[2] and Dr.A.Govardhan[3]

[1] Department of Computer Science and Engineering, Geethanjali College of Engineering and Technology Hyderabad, Andhra Pradesh 500007, India

[2] Department of Computer Science and Engineering, SLC'S IET Hyderabad, Andhra Pradesh 500028, India

[3] Department of Computer Science and Engineering, Jawaharlal Nehru Technological University, Jagtial, Andhra Pradesh 505327, India

Abstract

An emerging challenge for data mining is the problem of mining richly structured datasets, where the objects are linked in some way. Many real-world datasets describe a variety of object types linked via multiple types of relations. These links provide additional context that can be helpful for many data mining tasks. Links among the objects may demonstrate certain patterns, which can be helpful for many data mining tasks and are usually hard to capture with traditional statistical models. Recently there has been a surge of interest in this area, fuelled largely by interest in web and hypertext mining in personalization.

Keywords: *Link mining, Clustering, categorizer, Indexer, Personalization.*

1. Introduction

There are different traditional data mining tasks such as association rule mining, market basket analysis and cluster analysis commonly attempt to find patterns in a dataset characterized by a collection of independent instances of a single relation. This is consistent with the classical statistical inference problem of trying to identify a model given a random sample from a common underlying distribution. Data which is storing in data warehouse contains heterogeneous data. A key challenge for data mining is tackling the problem of mining richly structured, heterogeneous datasets. These datasets are typically multi-relational; they may be described by a relational database, a semi-structured representations such as XML, or using relational or first-order logic. However, the key commonalities are that the domain consists of a variety of object types and objects can be linked in some manner. In this case, the instances in our dataset are linked in some way, either by an explicit link, such as a URL, or by a constructed link, such as a join operation between tables stored in a database.

2. Link Mining

Link mining is a newly emerging research area that is at the intersection of the work in link analysis [4; 5], hypertext and web mining [3]. Links have more generically relationships, among data instances are ubiquitous. These links often exhibit patterns that can indicate properties of the data instances such as the importance, rank, or category of the object. In some cases, not all links will be observed. Therefore, we may be interested in predicting the existence of links between instances. In other domains, where the links are evolving over time, our goal may be to predict whether a link will exist in the future, given the previously observed links. By taking links into account, more complex patterns arise as well. This leads to other challenges focused on discovering substructures, such as communities, groups, or common sub graphs. Traditional data mining algorithms such as association rule mining, market basket analysis, and cluster analysis commonly attempt to find patterns in a dataset characterized by a collection of independent instances of a single relation. This is consistent with the classical statistical inference problem of trying to identify a model given a independent, identically distributed (IID) sample. One can think of this process as learning a model for the node attributes of a homogeneous graph while ignoring the links between the nodes. Link mining tasks are broadly categorized into following tasks. They are

1. Object-Related Tasks
    (a) Link-Based Object Ranking
    (b) Link-Based Object Classification

(c) Object Clustering (Group Detection)
(d) Object Identification (Entity Resolution)
2. Link-Related Tasks
(a) Link Prediction
3. Graph-Related Tasks
(a) Sub graph Discovery
(b) Graph Classification
(c) Generative Models for Graphs

In personalized web mining, which is considering different types of categorical domains. Personalization can be achieved through link mining by dynamically constructing user-profiles [1].

3. Proposed Algorithm to Construct Dynamic user Profiles using Link Mining

In this algorithm, thesis is concentrating on the hyperlinks of the web and different attributes of the link such as time spent on each link, link type, link category etc.

Process
Step 1: Get the log file (From server)
Step 2: Extract the Links and time spent browsing the links
Step 3: Check for the relevance of the document (t is time of viewing the document, th threshold time, If t > th the document is relevant)
Step 4: Categorize the documents that are found relevant
Step 5: Increase the score for the categories in Profile

In the above algorithm active time and passive time should be considered. If any user is visiting a link then finding about how much time he/she is actively reviewing that link (active time) or just visited that link and if he/she disconnects or away out of the desk should be considered (passive) to calculate the threshold time.

4. Architecture for constructing user Profiles

A closely related line of work is hypertext and web page classification. This work has its roots in the information retrieval (IR) community. A hypertext collection has a rich structure that should be exploited to improve classification accuracy. In addition to words, hypertext has both incoming and outgoing links. Traditional IR document models do not make full use of the link structure of hypertext. In the web page classification problem, the web is viewed as a large directed graph. Our objective is to label the category of a web page, based on features of the current page and features of linked neighbours. With the use of linkage information, such as anchor text and neighbouring text around each incoming link, better categorization results can be achieved.

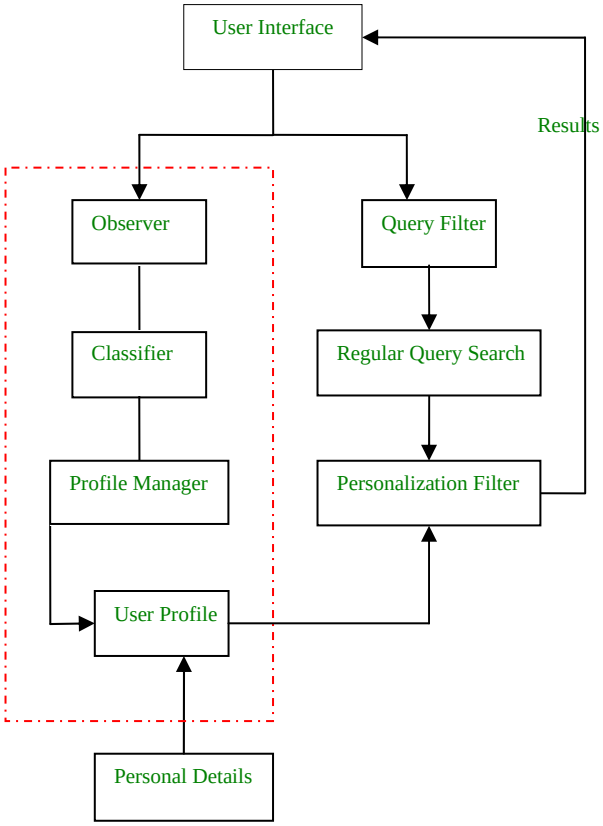

Fig. 1 Architecture for creating dynamic user profiles.

In the given architecture (figure 1) there are different modules with the following functionality.
Observer deals with the details of obtaining the information we need to build the user profile. Like,
Input Query
Links the user is visiting
Time the user is spending on each link
Length of the content of the link
Type of the link (PDF, doc, txt, images, audio, video)

Classifier deals with the methodology for query classification, where our aim is to classify queries onto a commercial taxonomy of Web queries with approximately 6000 nodes. Given such classifications, one can directly use them to provide better search results as well as more focused ads. The problem of query classification is extremely difficult owing to the brevity of queries. Observe, however, that in many cases a human looking at the search query and the search results does remarkably well in making sense of it. For instance, in the example

above, sending the query “SD450" to a Web search engine brings pages about Canon cameras, while “nc4200" brings pages about Compaq laptops, hence to a human the intent is quite clear. Of course, the sheer volume of search queries does not lend itself to human supervision, and therefore we need alternate sources of knowledge about the world. Search engines index colossal amounts of information, and as such can be viewed as large repositories of knowledge. Following the heuristic described above, we propose to use the search results themselves to gain additional insights for query interpretation. To this end, we employ the pseudo relevance feedback paradigm, and assume the top search results to be relevant to the query. Certainly, not all results are equally relevant, and thus we use elaborate voting schemes in order to obtain reliable knowledge about the query. For the purpose of this study we first dispatch the given query to a general Web search engine, and collect a number of the highest-scoring URLs. We crawl the Web pages pointed to by these URLs, and classify these pages. Finally, we use these result-page classifications to classify the original query. Our empirical evaluation confirms that using Web search results in this manner yields substantial improvements in the accuracy of query classification. Note that in a practical implementation of our methodology within a commercial search engine, all indexed pages can be pre-classified using the normal text- processing and indexing pipeline. Thus, at run-time we only need to run the voting procedure, without doing any crawling or classification. This additional overhead is minimal, and therefore the use of search results to improve query classification is entirely feasible at run-time.

Another important aspect of our work lies in the choice of queries. The volume of queries in today's search engines follows the familiar power law, where a few queries appear very often while most queries appear only a few times. While individual queries in this long tail are rare, together they account for a considerable mass of all searches. Furthermore, the aggregate volumes of such queries provide a substantial opportunity for income through on-line advertising. For frequent queries, searching and advertising platforms can be trained to provide good results, including auxiliary data such as maps, shortcuts to related structured information, successful ads, and so on. “Tail" queries, however, simply do not have enough occurrences to allow statistical learning on a per-query basis. Therefore, we need to aggregate such queries in some way, and to reason at the level of aggregated query clusters. A natural choice for such aggregation is to classify the queries into a topical taxonomy. Knowing which taxonomy nodes are most relevant to the given query will aid us to provide the same type of support for rare queries as for frequent queries. Consequently, in this work we focus on the classification of rare queries, whose correct classification is likely to be particularly beneficial.

Profile Manager takes Classifier’s inputs and uses them to refine/adapt/update the user profile.

User Profile contains the details and interests of the user.

5. Conclusions

Query classification is an important information retrieval task. Accurate classification of search queries can potentially be useful in a number of higher-level tasks such as Web search and ad matching. Since search queries are usually short, by themselves they usually carry insufficient information for adequate classification accuracy. To address this problem, we proposed a methodology for using search results as a source of external knowledge. To this end, we send the query to a search engine, and assume that a plurality of the highest-ranking search results is relevant to the query. Classifying these results and then allows us to classify the original query with substantially higher accuracy. There has been a growing interest in learning from linked data between objects. Tasks include hypertext classification, segmentation, information extraction, searching and information retrieval, discovery of authorities and link discovery. Domains include the world-wide web, bibliographic citations, criminology and bio-informatics, to name just a few. Learning tasks range from predictive tasks, such as classification, to descriptive tasks, such as the discovery of frequently occurring sub-patterns. There are other different data mining challenges in link mining such as identify of the, Link discovery, common relational patterns, where these topics lie in research area.

References

[1] Personalized Web Search by Mapping User Queries to Categories- Fang Liu Clement Yu Weiyi Meng Department of Computer Science, Department of Computer Science, Department of Computer Science, University of Illinois at Chicago University of Illinois at Chicago SUNY at Binghamton Chicago.

[2] D. Jensen. Statistical challenges to inductive inference in linked data. In Seventh International Workshop on Artificial Intelligence and Statistics, 1999 S. Zhang, C. Zhu, J. K. O. Sin, and P. K. T. Mok, “A novel ultrathin elevated channel low-temperature poly-Si TFT,” IEEE Electron Device Lett., vol. 20, pp. 569–571, Nov. 1999.

[3] S. Chakrabarti. Mining the Web. Morgan Kaufman, 2002.

[4] Personalized Web Search by Mapping User Queries to Categories- Fang Liu Clement Yu Weiyi Meng Department of Computer Science, Department of Computer Science, Department of Computer Science, University of Illinois at Chicago University of Illinois at Chicago SUNY at Binghamton Chicago.

[5] D. Jensen. Statistical challenges to inductive inference in linked data. In Seventh International Workshop on Artificial Intelligence and Statistics, 1999S. Zhang, C. Zhu, J. K. O. Sin, and P. K. T. Mok, "A novel ultrathin elevated channel low-temperature poly-Si TFT," IEEE Electron Device Lett., vol. 20, pp. 569–571, Nov. 1999.
[6] Beitzel, S., Jensen, E., Chowdhury, A., Grossman, D., and Frieder, O. 2004. Hourly analysis of a very large topically categorized web query log. In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM Press, Sheffield, United Kingdom, 321-328.
[7] Broder, A., P., Fontoura, M., Gabrilovich, E., Josifovski, V., and Reidel, L.2008. Search advertising using Web relevance feedback. In CIKM'08.
[8] Gabrilovich, E. and Markovitch,S.2007. Harnessing the expertise of 70,000 human editors: Knowledge-based feature generation for text categorization. Journal of Machine Learning Research 8,2297-2345.

K.Srinivas is a Ph.D. student and received Mater of Technology in Computer Science and Engineering from Jawaharlal Nehru Technological University in 2005 and Bachelor of Engineering in computer science and Engineering from Swami Ramananda Teerth Marathwada University in 2000. He is working as Associate Professor in Geethanjali College of Engineering and Technology and worked as Assistant Professor in Arkay College of Engineering and Technology affiliated to Jawaharlal Nehru Technological University. His main research interests include Data Mining and Information Retrieval.

L.Kiran Kumar Reddy received M.Tech. in Computer Science and Engineering in 2006 and B.E. in Computer Science and Engineering from Swami Ramananda Teerth Marathwada University in 2000. He is working as Associate Professor in Satyam Learning Campus's Institute of Engineering and Technology and worked as Assistant Professor in Arkay College of Engineering and Technology. He is doing research in Data Mining and Information Retrieval.

Dr.A.Govardhan received Ph.D. degree in Computer Science and Engineering from Jawaharlal Nehru Technological University in 2003 M.Tech. from Jawaharlal Nehru University in 1994 and B.E. in from Osmania University in 1992. He is Working as a Principal of Jawaharlal Nehru Technological University, Jagitial. He has Published around 50 papers in various national and international Journals/conferences. His research of interest includes Data Mining, Information Retrieval, Search Engines.