Data Mining in Web Applications for Business Intelligence Operations

[1]V Veera Ankalu, [2]K Ravikiran, [3]K Rama Rao
[1]Dept. of CSE, Jayawant College of Engg. & Mgmt., Maharashtra, India
[2,3]Dept. of IT, Sri Sunflower College of Engg & Tech., Lankapalli, AP, India

Abstract
The exponential explosion of various contents generated on the Web, Recommendation techniques have become increasingly indispensable. Innumerable different kinds of recommendations are made on the Web every day, including movies, music, images, books recommendations, query suggestions, tags recommendations, etc. Over the last decade, there has been a paradigm shift in business computing with the emphasis moving from data collection to knowledge extraction. Central to this shift has been the explosive growth of the World Wide Web, which has enabled myriad technologies, such as Web services and enterprise server applications. These advances have improved data collection frameworks and resulted in new techniques for knowledge extraction from large databases. A popular and successful technique which has showed much promise is Web mining. Web mining is essentially data mining for Web data, thus enabling businesses to turn their vast repositories of transactional and Website usage data into actionable knowledge that is useful at every level of the enterprise – not just the front-end of an online store. To this end, the chapter provides an introduction to the field of Web mining and examines existing as well as potential Web mining applications applicable for different business function, like marketing, human resources, and fiscal administration. Suggestions for improving information technology infrastructure are made, which can help businesses interested in Web mining hit the ground running.

Keywords
Web Mining, Collection Framework, Business Intelligence, Graph Mining

I. Introduction

Web Mining can be broadly divided into three distinct categories, according to the kinds of data to be mined:

A. Web Content Mining

Web Content Mining is the process of extracting useful information from the contents of Web documents. Content data corresponds to the collection of facts a Web page was designed to convey to the users. It may consist of text, images, audio, video, or structured records such as lists and tables. Text mining and its application to Web content has been the most widely researched. Some of the research issues addressed in text mining are, topic discovery, extracting association patterns, clustering of web documents and classification of Web Pages. Research activities in this field also involve using techniques from other disciplines such as Information Retrieval (IR) and Natural Language Processing (NLP). While there exists a significant body of work in extracting knowledge from images - in the fields of image processing and computer vision - the application of these techniques to Web content mining has not been very rapid.

B. Web Structure Mining

The structure of a typical Web graph consists of Web pages as nodes, and hyperlinks as edges connecting between two related pages. Web Structure Mining can be regarded as the process of discovering structure information from the Web.
This type of mining can be further divided into two kinds based on the kind of structural data used.

1. Hyperlinks

A Hyperlink is a structural unit that connects a Web page to different location, either within the same Web page or to a different Web page. A hyperlink that connects to a different part of the same page is called an Intra-Document Hyperlink, and a hyperlink that connects two different pages is called an Inter-Document Hyperlink. There has been a significant body of work on hyperlink analysis, of which provides an up-to-date survey.

2. Document Structure

In addition, the content within a Web page can also be organized in a tree-structured format, based on the various HTML and XML tags within the page. Mining efforts here have focused on automatically extracting document object model (DOM) structures out of documents.

C. Web Usage Mining

Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data, in order to understand and better serve the needs of Web-based applications. Usage data captures the identity or origin of Web users along with their browsing behavior at a Web site. Web usage mining itself can be classified further depending on the kind of usage data considered:

D. Web Server Data

They correspond to the user logs that are collected at Web server. Some of the typical data collected at a Web server include IP addresses, page references, and access time of the users.

E. Application Server Data

Commercial application servers, e.g. Weblogic[BEA], BroadVision [BV], StoryServer [VIGN], etc. have significant features in the framework to enable E-commerce applications to be built on top of them with little effort. A key feature is the ability to track various kinds of business events and log them in application server logs.

F. Application Level Data

Finally, new kinds of events can always be defined in an application, and logging can be turned on for them – generating histories of these specially defined events.
The usage data can also be split into three different kinds on the basis of the source of its collection: on the server side, the client side, and the proxy side. The key issue is that on the server side there is an aggregate picture of the usage of a service by all users, while on the client side there is complete picture of usage of all services by a particular client, with the proxy side being somewhere in the middle.

II. Notations

- The Notations are as follows:
- G = (V, E)
- V: set of N vertices
- E _ VxV : set of edges directed or undirected
- N(u) = {v|(u, v) 2 E}: neighbors of u
- d(u) = |N(u)|: degree of u

Examples

- Webgraph: Web pages and hyperlinks.
- Social networks: Members and acquaintance relationships.
- Internet: ASs and peering relationships;
- P2P networks: Peers and transactions.

A. Degree Distribution

- Ck = |{u : d(u) = k}|: number of vertices with degree
- (indegree/outdegree) k. We often observe power law distribution:
- $Ck = ck^{-\alpha}$,
- with $\alpha \geq 1$, or
- $\ln Ck = -\alpha (\ln c + \ln k)$.
- Plotting lnCk versus lnk gives a straight line with slope $-\alpha$.
- Heavy tail distribution: there is a non-negligible fraction of nodes that have very high degree.

III. Developer Duplication Reduction

Many software businesses, both large and small, maintain one or more internal application development units. Thus, at any given time, there may be hundreds, if not thousands, of projects being developed, deployed, and maintained concurrently. Due to overlapping business processes (i.e. human resources and fiscal administration) and multiple project development groups, duplication of source code often occurs (Rajapakse and Jarzabek, 2005) and (Kapser and Godfrey, 2003). Given the non-trivial cost of application development, mitigating such duplication is critical. Source code consistency is also an issue, e.g. to prevent a case where only one of two duplicate segments is updated to address a bug and/or feature addition.

Turnkey solutions for source code duplication are already available, but they suffer from two major problems:

They are not able to address code which is functionally similar, but syntactically different.

They only detect duplication after it has already occurred.

Figure 1 gives an overview of a possible approach for identifying potential duplication among multiple projects.

First, the project Web pages and documents must be extracted from the intranet. Next, each document is split into fragments using common separators (periods, commas, bullet points, new lines, etc). These fragments form the most basic element of comparison – the smallest entity capable of expressing a single thought. Using clustering techniques, these fragments can then be grouped into collections of similar fragments. When two or more fragments are part of the same collection, but come from different projects, potential duplication has been identified. These fragments may then be red-flagged and brought to the attention of affected project managers.

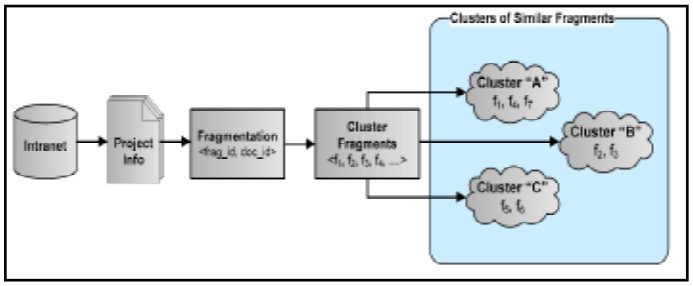


Fig. 1: Duplication Candidate Process Overview

Expert Driven Recommendations for Customer Assistance Most recommender systems used in business today are product-focused, where recommendations made to a customer are typically a function of his/her interests in products (based on his/her browsing history) and that of other similar customers. However, in many cases, recommendations must be made without knowledge about a customer's preferences, like is customer service call centres. In such cases, call centre employees leverage their domain knowledge in order to help align customer inquiries with appropriate answers. Here, a customer may be wrong, which is often observed when domain experts are asked questions by non-experts.

Many businesses must maintain large customer service call centres, especially in retail-based operations, in order to address this need. However, advances in Web-based recommender systems may enable to improve call center capacity by offering expert-based recommendations online (Delong et al, 2005).

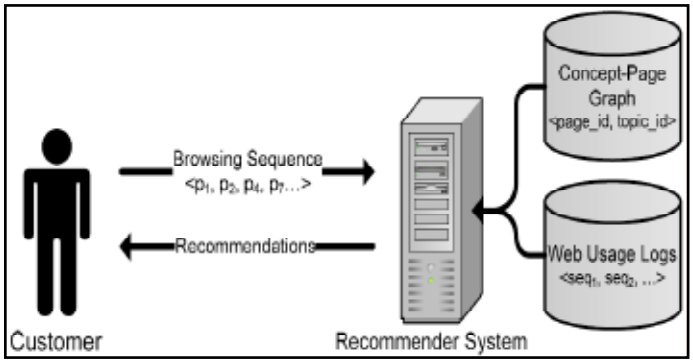


Fig. 2: Overview of Expert-Driven Customer Assistance Recommendations

IV. Business Process Mining

Business process mining, also called workflow mining, reveals how existing processes work and, thus providing considerable ROI (PMR). Business process mining is the task of extracting useful information from business event logs collected by Workflow Management Systems such as IBM's WebSphere and SAP R/3. ProM, EMiT and Thumb are some examples of business process mining tools. An example process mining framework is shown in fig. 3.

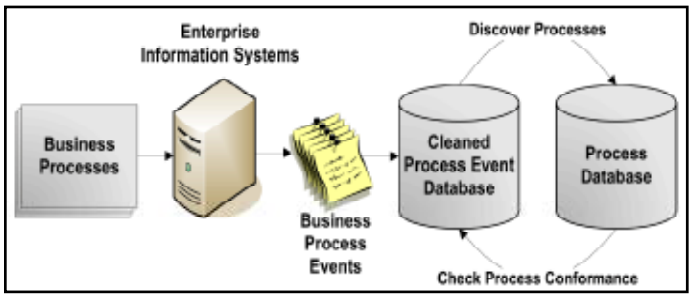


Fig. 3: An Example Process Mining Framework

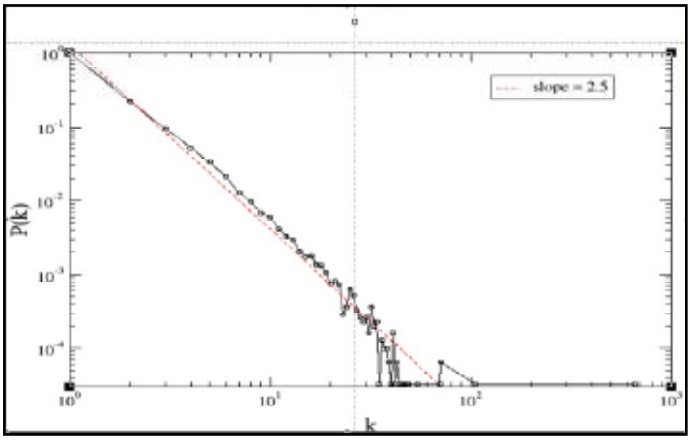
Business transaction logs obtained from Enterprise Information systems are transformed into XML format.

These event logs are then cleaned and time-ordering of the processes is inferred. Using these, business process models are built and business rules are constructed. Finally, these process models are converted to a Petri-Net for analysis. The process model discovered can also be checked for conformance with previously discovered models. We can also use an anomaly detection system to identify deviation in business process behaviour.

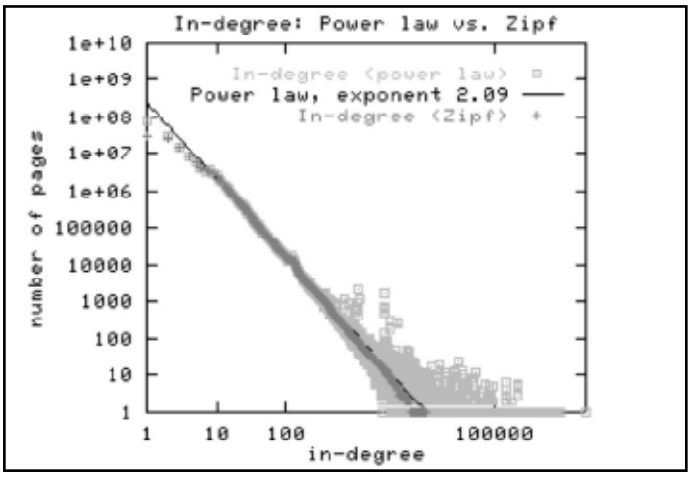
V. Graph Mining

In this we a can see the different types graph show the requirements analysis in real time.

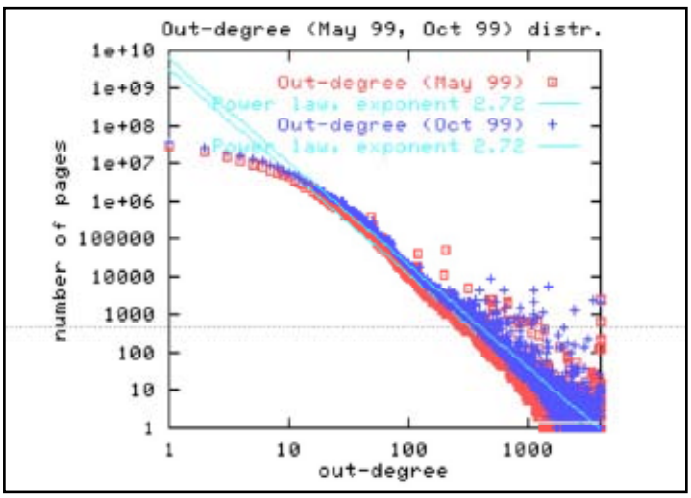
Internet graph [Faloutsos et al.,1999]



Webgraph Indegree[Broder et al., 2000, Donato et al., 2007]



Webgraph outdegree[Broder et al., 2000, Donato et al., 2007]



Other degree-related metrics

- Edge-reciprocity: Percentage of links that are reciprocal
- Degree/ average degree of neighbors. Assortative behavior if highly linked vertices are linked to vertices of high degree, disassortative otherwise.
- Average indegree of out neighbors.
- Average outdegree of in neighbors.

VI. Conclusion

This work proposes a scalable approximate data mining mechanism to compress the web graph for link servers, which incorporates the formation of global patterns. This chapter examines how technology, such as Web mining, can aid businesses in gaining an extra information and intelligence. We provide an introduction to Web mining and the various techniques associated with it. We briefly update the reader with state-of-art research in this area. Later, we show how these class of techniques can be effectively used to aid various business functions. We provide example applications to illustrate the use of Web mining techniques to aid in certain areas of business functions. These examples provide the evidence of success and the potential of Web mining for business intelligence. Finally, we point out the gaps in existing technologies and certain future directions that should be of interest to the business community at large. Directions for future work include a parallel formulation, and a mechanism to grow larger communities from the discovered seed patterns.

References

[1] Rajapakse, D. C. Jarzabek, S.,"An Investigation of Cloning in Web Applications", Fifth International Conference on Web Engineering, Sydney, Australia, July 27-29, 2005.

[2] Kapser, C., Godfrey, M. W.,"Toward a Taxonomy of Clones in Source Code: A Case Study", International Workshop on Evolution of Large-scale Industrial Software Applications, Amsterdam, The Netherlands, 2003.

[3] Prasanna Desikan, Colin DeLong, Sandeep Mane, Kalyan Beemanapalli, Kuo-Wei Hsu, Prasad Sriram, Jaideep Srivastava, Vamsee Venuturumilli,"Web Mining for Business Computing".

[4] Getoor, L.,"Link Mining: A New Data Mining Challenge", SIGKDD Explorations, 4(2), 2003.

[5] Kapser, C., Godfrey, M. W.,"Toward a Taxonomy of Clones in Source Code: A Case Study", International Workshop on Evolution of Large-scale Industrial Software Applications, Amsterdam, The Netherlands, 2003.

[6] Marc Najork, Microsoft Research,"Mining the Web Graph".

[7] Stefano Leonardi,"Graph Mining and its applications to Reputation Management in Networks", Sapienza University of Rome – Rome, Italy.

Vuyyuru Veera Ankalu received his B.TECH degree in CSE from Gudlavalleru Engineering College Krishna(DT), in 2006, the M.TECH. degree in CST from GITAM UNIVERSITY, VISAKHAPATNAM in 2010. At present, He is working as Asst.Professor in Jayawant College of Engg & Mgmt,Karad, Maharastra, India

Ravikiran Kolagani received the BTech degree from Newtons Institute Of Engg And Technology in 2009. M.TECH. degree in CSE from Sri Sunflower College Of Engg And Technology in 2012. Currently he is working as AN Asst.Professor in Sri Sunflower College Of Engg And Technology, Lankapalli, Andhra Pradesh, India.

Katru Rama Rao received the BTech degree IN CSIT from St Theressa Institute Of Engg And Technology in 2005. M.Tech. degree in CSE from SRI Sunflower College Of Engg And Technology in 2012. Currently he is working as AN Asst.Professor in SRI Sunflower College Of Engg And Technology, Lankapalli, Andhra Pradesh, India.