A Generalised Flow-Based Method For Analysis Of Implicit Relationships On Wikipedia

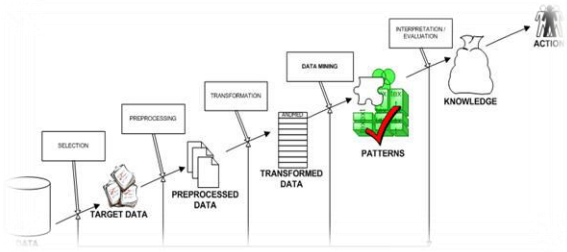
Durganath Rachamalla, Sri.G.Ramesh Kumar, Sri.V.Bhaskara Murthy
MCA Student, Assistant Professor, Associate Professor
Dept Of MCA
B.V.Raju College, Bhimavaram

ABSTRACT

We focus on measuring relationships between pairs of objects in Wikipedia whose pages can be regarded as individual objects. Two kinds of relationships between two objects exist: in Wikipedia, an explicit relationship is represented by a single link between the two pages for the objects, and an implicit relationship is represented by a link structure containing the two pages. Some of the previously proposed methods for measuring relationships are cohesion-based methods, which underestimate objects having high degrees, although such objects could be important in constituting relationships in Wikipedia. The other methods are inadequate for measuring implicit relationships because they use only one or two of the following three important factors: distance, connectivity, and co citation. We propose a new method using a generalized maximum flow which reflects all the three factors and does not underestimate objects having high degree. We confirm through experiments that our method can measure the strength of a relationship more appropriately than these previously proposed methods do. Another remarkable aspect of our method is mining elucidatory objects, that is, objects constituting a relationship. We explain that mining elucidatory objects would open a novel way to deeply understand a relationship.

I. INTRODUCTION

What is Data Mining?



Structure of Data Mining

Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

How Data Mining Works?

While large-scale information technology has been evolving separate transaction and analytical systems, data mining provides the link between the two. Data mining software analyzes relationships and patterns in stored transaction data based on open-ended user queries. Several types of analytical software are available: statistical, machine learning, and neural

networks. Generally, any of four types of relationships are sought:

- Classes: Stored data is used to locate data in predetermined groups. For example, a restaurant chain could mine customer purchase data to determine when customers visit and what they typically order. This information could be used to increase traffic by having daily specials.

- Clusters: Data items are grouped according to logical relationships or consumer preferences. For example, data can be mined to identify market segments or consumer affinities.

- Associations: Data can be mined to identify associations. The beer-diaper example is an example of associative mining.

- Sequential patterns: Data is mined to anticipate behavior patterns and trends. For example, an outdoor equipment retailer could predict the likelihood of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes.

Data mining consists of five major elements:

1) Extract, transform, and load transaction data onto the data warehouse system.
2) Store and manage the data in a multidimensional database system.
3) Provide data access to business analysts and information technology professionals.
4) Analyze the data by application software.
5) Present the data in a useful format, such as a graph or table.

Different levels of analysis are available:

- Artificial neural networks: Non-linear predictive models that learn through training and resemble biological neural networks in structure.

- Genetic algorithms: Optimization techniques that use process such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution.

- Decision trees: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID). CART and CHAID are decision tree techniques used for classification of a dataset. They provide a set of rules that you can apply to a new (unclassified) dataset to predict which records will have a given outcome. CART segments a dataset by creating 2-way splits while CHAID segments using chi square tests to create multi-way splits. CART typically requires less data preparation than CHAID.

- Nearest neighbor method: A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where k=1). Sometimes called the k-nearest neighbor technique.

- Rule induction: The extraction of useful if-then rules from data based on statistical significance.

- Data visualization: The visual interpretation of complex relationships in multidimensional data. Graphics tools are used to illustrate data relationships.

Characteristics of Data Mining:

- Large quantities of data: The volume of data so great it has to be analyzed by automated techniques e.g. satellite information, credit card transactions etc.
- Noisy, incomplete data: Imprecise data is the characteristic of all data collection.
- Complex data structure: conventional statistical analysis not possible
- Heterogeneous data stored in legacy systems

Benefits of Data Mining:

1) It's one of the most effective services that are available today. With the help of data mining, one can discover precious information about the customers and their behavior for a specific set of products and evaluate and analyze, store, mine and load data related to them
2) An analytical CRM model and strategic business related decisions can be made with the help of data mining as it helps in providing a complete synopsis of customers
3) An endless number of organizations have installed data mining projects and it has helped them see their own companies make an unprecedented improvement in their marketing strategies (Campaigns)
4) Data mining is generally used by organizations with a solid customer focus. For its flexible nature as far as applicability is concerned is being used vehemently in applications to foresee crucial data including industry analysis and consumer buying behaviors
5) Fast paced and prompt access to data along with economic processing techniques have made data mining one of the most suitable services that a company seek

Advantages of Data Mining:

1. Marketing / Retail:

Data mining helps marketing companies build models based on historical data to predict who will respond to the new marketing campaigns such as direct mail, online marketing campaign…etc. Through the results, marketers will have appropriate approach to sell profitable products to targeted customers.

Data mining brings a lot of benefits to retail companies in the same way as marketing. Through market basket analysis, a store can have an appropriate production arrangement in a way that customers can buy frequent buying products together with pleasant. In addition, it also helps the retail companies offer certain discounts for particular products that will attract more customers.

2. Finance / Banking

Data mining gives financial institutions information about loan information and credit reporting. By building a model from historical customer's data, the bank and financial institution can determine good and bad loans. In addition, data mining helps banks detect

fraudulent credit card transactions to protect credit card's owner.

3. Manufacturing

By applying data mining in operational engineering data, manufacturers can detect faulty equipments and determine optimal control parameters. For example semi-conductor manufacturers has a challenge that even the conditions of manufacturing environments at different wafer production plants are similar, the quality of wafer are lot the same and some for unknown reasons even has defects. Data mining has been applying to determine the ranges of control parameters that lead to the production of golden wafer. Then those optimal control parameters are used to manufacture wafers with desired quality.

4. Governments

Data mining helps government agency by digging and analyzing records of financial transaction to build patterns that can detect money laundering or criminal activities.

5. Law enforcement:

Data mining can aid law enforcers in identifying criminal suspects as well as apprehending these criminals by examining trends in location, crime type, habit, and other patterns of behaviors.

6. Researchers:

Data mining can assist researchers by speeding up their data analyzing process; thus, allowing those more time to work on other projects.

II. EXISTING SYSTEM

Several methods have been proposed for measuring the strength of a relationship between two objects on an information network (V, E), a directed graph where V is a set of objects; an edge (u, v)∈ E exists if and only if object u ∈ V has an explicit relationship to u ∈ V. We can define a Wikipedia information network whose vertices are pages of Wikipedia and whose edges are links between pages. Previously proposed methods then can be applied to Wikipedia by using a Wikipedia information network. The Concept of "cohesion," exists for measuring the strength of an implicit relationship. CFEC proposed by Koren et al. [1] and PFIBF proposed by Nakayama et al. is based on cohesion. We do not adopt the idea of cohesion based methods, because they always punish objects having high degrees although such objects could be important to some relationships in Wikipedia. Other previously proposed methods use only one or two of the three representative concepts for measuring a relationship: distance, connectivity, and cocitation, although all the concepts are important factors for implicit relationships. Using all the three concepts together would be appropriate for measuring an implicit relationship and mining elucidatory objects.

DISADVANTAGES OF EXISTING SYSTEM:

- It is difficult for the user to discover an implicit relationship and elucidatory objects without investigating a number of pages and links.
- Therefore, it is an interesting problem to measure and explain the strength of an implicit relationship between two objects in Wikipedia.

III. PROPOSED SYSTEM

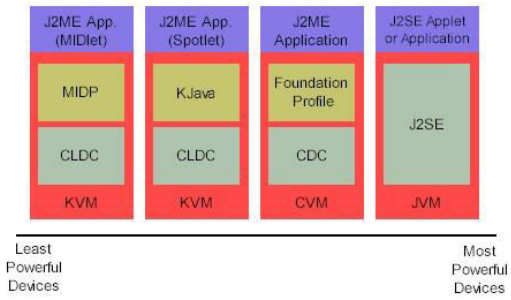
We propose a new method for measuring a relationship on Wikipedia by reflecting all the three concepts: distance, connectivity, and

cocitation. We measure relationships rather than similarities. As discussed in relationship is a more general concept than similarity. For example, it is hard to say petroleum is similar to USA, but a relationship exists between petroleum and the USA. Our method uses a "generalized maximum flow" on an information network to compute the strength of a relationship from object s to object t using the value of the flow whose source is s and destination is t. It introduces a gain for every edge on the network. The value of a flow sent along an edge is multiplied by the gain of the edge. Assignment of the gain to each edge is important for measuring a relationship using a generalized maximum flow. We propose a heuristic gain function utilizing the category structure in Wikipedia. We confirm through experiments that the gain function is sufficient to measure relationships appropriately.

ADVANTAGES OF PROPOSED SYSTEM:

- Compute the strength of the relationship between a source object and each of its destination objects, and rank the destination objects by the strength.
- Assignment of the gain to each edge is important for measuring a relationship using a generalized maximum flow.
- Experiments on Wikipedia showing that our method is the most appropriate one

IV. SYSTEM ARCHITECTURE



V. IMPLEMENTATION

MODULES:

- Link Analysis Module.
- Generalized Flow Based Module.
- Wikipedia Mining Module.
- Ranking Module.

MODULES DESCRIPTION:

Link Analysis Module:

Two kinds of relationships between two objects exist: in Wikipedia, an explicit relationship is represented by a single link between the two pages for the objects, and an implicit relationship is represented by a link structure containing the two pages. A user also might desire to discover a relationship between two objects. For example, a user might desire to know which countries are strongly related to petroleum, or to know why one country has a stronger relationship to petroleum than another country. Typical keyword search engines can neither measure nor explain the strength of a relationship. The main issue for measuring relationships arises from the fact that two kinds of relationships: "explicit relationships" and "implicit relationships." In Wikipedia, an explicit relationship is represented by a link. An implicit relationship is represented by multiple links and pages. For example, an implicit relationship between petroleum and the USA might be represented by links.

Generalized Flow Based Module:

The three concepts, distance, connectivity, and cocitation, are important concepts for measuring relationships; cohesion-based methods underestimate popular objects, although popular objects might be important for relationships in Wikipedia. Our method can mine elucidatory objects constituting a relationship by outputting paths contributing to the generalized maximum flow, that is, paths along which a large amount of flow is sent. We propose a generalized maximum flow-based method which reflects all the three concepts and does not under estimates popular objects, in order to measure relationships on Wikipedia appropriately.

Wikipedia Mining Module:

Searching webpages containing a keyword has grown in this decade, while knowledge search has recently been researched to obtain knowledge of a single object and relationships between multiple objects, such as humans, places or events. Searching knowledge of objects using Wikipedia is one of the hottest topics in the field of knowledge search. In Wikipedia, the knowledge of an object is gathered in a single page updated constantly by a number of volunteers. Wikipedia also covers objects in a number of categories, such as people, science, geography, politic, and history. Therefore, searching Wikipedia is usually a better choice for a user to obtain knowledge of a single object than typical search engines.

Ranking Module:

We propose a new method for measuring the strength of a relationship using the generalized maximum flow. In addition we propose a ranking module, where the A good evaluation of methods measuring relationships always requires human subjects ranking.

VI. CONCLUSION

We have proposed a new method of measuring the strength of a relationship between two objects on Wikipedia. By using a generalized maximum flow, the three representative concepts, distance, connectivity, and cocitation, can be reflected in our method. Furthermore, our method does not underestimate objects having high degrees. We have ascertained that we can obtain a fairly reasonable ranking according to the strength of relationships by our method compared with those by GSD [7], PFIBF [3], [2], CFEC [1], and THT [12]. Particularly, our method is the only choice for measuring 3-hop implicit relationships. We have also confirmed that elucidatory objects are helpful to deeply understand a relationship. Some future challenges remain. We are also interested in seeking possibilities of the elucidatory objects constituting a relationship mined by our method. We plan to quantitatively evaluate the elucidatory objects. We are developing a tool for deeply understanding relationships by utilizing elucidatory objects.

REFERENCES

[1] Y. Koren, S.C. North, and C. Volinsky, “Measuring and Extracting Proximity in Networks,” Proc. 12th ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining, pp. 245-255, 2006.

[2] M. Ito, K. Nakayama, T. Hara, and S. Nishio, “Association Thesaurus Construction Methods Based on Link Co-Occurrence Analysis for Wikipedia,” Proc. 17th ACM Conf. Information and Knowledge Management (CIKM), pp. 817-826, 2008.

[3] K. Nakayama, T. Hara, and S. Nishio, “Wikipedia Mining for an Association Web Thesaurus Construction,” Proc. Eighth Int’l Conf. Web Information Systems Eng. (WISE), pp. 322-334, 2007.

[4] J. Gracia and E. Mena, “Web-Based Measure of Semantic Relatedness,” Proc. Ninth Int’l

Conf. Web Information Systems Eng. (WISE), pp. 136-150, 2008.
[5] R.K. Ahuja, T.L. Magnanti, and J.B. Orlin, Network Flows: Theory, Algorithms, and Applications. Prentice Hall, 1993.
[6] K.D. Wayne, "Generalized Maximum Flow Algorithm," PhD dissertation, Cornell Univ., New York, Jan. 1999.
[7] R.L. Cilibrasi and P.M.B. Vita´nyi, "The Google Similarity Distance," IEEE Trans. Knowledge and Data Eng., vol. 19, no. 3, pp. 370-383, Mar. 2007.
[8] G. Kasneci, F.M. Suchanek, G. Ifrim, M. Ramanath, and G. Weikum, "Naga: Searching and Ranking Knowledge," Proc. IEEE 24th Int'l Conf. Data Eng. (ICDE), pp. 953-962, 2008. [9] F.M. Suchanek, G. Kasneci, and G. Weikum, "Yago: A Core of Semantic Knowledge," Proc. 16th Int'l Conf. World wide Web Conf. (WWW), pp. 697-706, 2007.
[10] "The Erdo¨ s Number Project," http://www.oakland.edu/enp/, 2012.
[11] M. Yazdani and A. Popescu-Belis, "A Random Walk Framework to Compute Textual Semantic Similarity: A Unified Model for Three Benchmark Tasks," Proc. IEEE Fourth Int'l Conf. Semantic Computing (ICSC), pp. 424-429, 2010.
[12] P. Sarkar and A.W. Moore, "A Tractable Approach to Finding Closest Truncated-Commute-Time Neighbors in Large Graphs," Proc. 23rd Conf. Uncertainty in Artificial Intelligence (UAI), 2007.
[13] W. Lu, J. Janssen, E. Milios, N. Japkowicz, and Y. Zhang, "Node Similarity in the Citation Graph," Knowledge and Information Systems, vol. 11, no. 1, pp. 105-129, 2006.
[14] H.D. White and B.C. Griffith, "Author Cocitation: A Literature Measure of Intellectual Structure," J. Am. Soc. Information Science and Technology, vol. 32, no. 3, pp. 163-171, May 1981.
[15] D. Milne and I.H. Witten, "An Effective, Low-Cost Measure of Semantic Relatedness Obtained from Wikipedia Links," Proc. AAAI Workshop Wikipedia and Artificial Intelligence: An Evolving Synergy, 2008.