Flow-Based Methodology For Analysis Of Links On Wikipedia

PALAGIRI MOHAMMAD NOUSHAD, P SHABANA M-TECH

# Student of M.Tech., Srinivasa Ramanujan Institute Of Technology And Sciences, Rotarypuram.
# Assistant Professor, Srinivasa Ramanujan Institute Of Technology And Sciences, Rotarypuram.

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.

Keywords: Link analysis, generalized flow, Wikipedia mining, relationship.

I. INTRODUCTION

Searching web pages 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.

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 exist: “explicit relationships” and “implicit relationships.” In Wikipedia, an explicit relationship is represented by a link. For example, an explicit relationship between petroleum and Gulf of Mexico might be represented by a link from page “Petroleum” to page “Gulf of Mexico.” A user could understand its meaning by reading the text “Oil filed in Gulf of Mexico is a major petroleum producer” surrounding the anchor text “Gulf of Mexico” on page “Petroleum.” 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 and pages depicted in Fig. 1. For an implicit relationship between two objects, the objects, except the two objects, constituting the relationship is named
elucidatory objects because such objects enable us to explain the relationship. For the example described above, “Gulf of Mexico” is one of the elucidatory objects. The user can understand an explicit relationship between two objects easily by reading the pages for the two objects in Wikipedia. By contrast, 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.

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 [4], 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” [5], [6] 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.

II. RELATED WORK

“The Erdő’s number used by mathematicians is based on distance and coauthorships. The legendary mathematician Paul Erdő’s has a number 0, and the people who cowrote a paper with Erdő’s have a number 1; the people who cowrote a paper with a person with a number 1 have a number 2, and so on. The Erdő’s number is the distance, or the length of the shortest path, from a person to Erdő’s on an information network whose edge represents coauthorship; a shorter path represents a stronger relationship. However, the Erdő’s number is inadequate to represent the implicit relationship between a person and Erdő’s because the number does not estimate the connectivity between them. The hitting time [11], [12] from vertex s to vertex t is defined as the expected number of steps in a random walk starting from s before t is visited for the first time. Actually, the hitting time from s to t in a network represents the average length of all the paths connecting s and t. Sarkar and Moore [12] proposed “Truncated Hitting Time” (THT) to compute the average length of paths connecting two vertices whose length are at most Lmax only. A smaller distance represents a larger similarity. THT does not estimate the connectivity between two vertices. For example, suppose only m _ 1 vertex disjoint paths of length k connect s to t. THT computes the distance from s to t to be k for any m _ 1. We compare our method with THT through experiments in. The connectivity [5], more precisely the vertex connectivity, from vertex s to vertex t on a network is the minimum number of vertices such that no path exists from s to t if the vertices are removed. s has a strong relationship to t if the connectivity from s to t is large. The connectivity from s to t is equal to the value of a maximum flow from s to t, where every edge and vertex has capacity 1. However, the distance cannot be estimated by the maximum flow because the amount of a flow along a path is independent of the path length. Lu et al. [13] proposed a method for computing the strength of a relationship using a maximum flow. They tried to estimate the distance between two objects using a maximum flow by setting edge capacities. However, the value of a maximum flow does not necessarily decrease by setting only capacities even if the
distance becomes larger. Therefore, their method cannot estimate the distance successfully by the value of the maximum flow. Instead of setting capacities, we use a generalized maximum flow by setting every gain to a value less than one. Therefore, the value of a maximum flow in our method decreases if the distance becomes longer.

Cocitation-based methods assume that two objects have a strong relationship if the number of objects linked by both the two objects is large [14]. On the other hand, cooccurrence is a concept by which the strength is represented by the number of objects linking to both objects. The “Google Similarity Distance” proposed by Cilibrasi and Vitányi [7] can be regarded as a co-occurrence based method; it measures the strength of a relationship between two words by counting of webpages containing both words. That is, it implicitly regards the webpages as the objects linking to the two objects representing the two words. In an information network, an object linked by both objects becomes an object linking to the both if the direction of every edge is reversed. Therefore, co-occurrence can be regarded as the reverse of the cocitation. We then include cooccurrence-based methods among co-citation-based methods in this paper. Milne and Witten [15] also proposed methods measuring relationships between objects in Wikipedia using Wikipedia links based on cocitation. Cocitation-based methods cannot deal with a typical implicit relationship, such as “person w is regarded as a friend by person v who is regarded as a friend by person u.” This relationship is represented by the path formed by two edges δu; vP and δv; wP. In contrast, cocitation-based methods can deal with two edges going into the same vertex, such as edges δu; vP and δw; vP. Therefore, cocitation-based methods are inadequate for measuring an implicit relationship. Furthermore, cocitation-based methods cannot deal with 3-hop implicit relationships defined in Section 1 because these methods estimate only relationships represented by paths formed by two edges, as explained above. SimRank, proposed by Jeh and Widom [16], is an extension of cocitation-based methods. SimRank employs recursive computation of cocited objects, therefore it can deal with a path whose length is longer than two, although it cannot deal with a typical implicit relationship “a friend of a friend” similarly to cocitation-based methods. If we define all edges as bidirectional, then SimRank could measure the typical implicit relationship. However, we observed that SimRank computes the strength of the relationship represented by a path constituted by an odd number of edges to be 0, even if all edges are bidirectional. For example, SimRank computes the strength of the relationship between u and w to be 0 if the relationship is represented by path δu; wP or δu; v0; v1; wP. Such paths abound in the Wikipedia information network. Therefore, SimRank is inappropriate for measuring relationships on Wikipedia.

Cohesion

In the field of social network analysis, cohesion-based methods are known to measure the strength of a relationship by counting all paths between two objects. The original cohesion was proposed by Hubbell, Katz, Wasserman and K. Faust. It has a property that its value greatly increases if a popular object, an object linked from or to many objects, exists. As pointed out in other researches, this property is a defect for measuring the strength of a relationship. Several cohesion-based methods, such as PFIBF and CFEC explained below, were proposed to dissolve this property.

Sequential Pattern mining with the Integration of Fuzzy Logic and Graph Search Techniques

In this section, the researcher proposes a mining sequential pattern with the integration of fuzzy logic and graph search techniques. The method includes three points: (1) applies with sequential patterns (2) define the interval extended sequence using fuzzy sets, (3) searches sequential patterns using graph search techniques.
III. PROPOSED MODEL

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 the previous section, 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.

The main contributions of this paper are as follows:
1. A detailed and methodical survey of related work for measuring relationships or similarities
2. A new method using generalized maximum flow for measuring the strength of a relationship between two objects on Wikipedia, which reflects the three concepts: distance, connectivity, and co-citation
3. Experiments on Wikipedia showing that our method is the most appropriate one
4. Case studies of mining elucidatory objects for deeply understanding a relationship

IV. ARCHITECTURE

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 $(u; v) \in E$ exists if and only if object $u \in V$ has an explicit relationship to $v \in V$. 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. A concept “cohesion,” exists for measuring the strength of an implicit relationship. CFEC proposed by Koren et al. [1] and PFIBF proposed by Nakayama et al. [2], [3] are 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, as we will explain in Section 2.2. 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.
VI. CONCLUSION

The reductionist approach of viewing data mining in terms of statistics has advantages of the strong theoretical background and easy-formulated problems. The data compression and constructive induction approaches have relatively strong analytical background, as well as connections to the philosophy of science. In addition to the just mentioned frameworks an interesting solution is proposed in the microeconomic view on data mining, introduced by Kleinberg (1998), where a utility function is constructed and trying to be maximized. The data mining tasks concerning processes like clustering, regression, and classification fit easily into these approaches.

The inductive databases framework suggests architecture for data mining systems and allows to view data mining as a process. Association rules and other simple pattern formalisms can be described by this approach. However, for example, clustering is harder to describe in a useful way within the inductive databases framework.

In one way or another, we can easily see the exploratory nature of the frameworks for the data-mining field. Different frameworks account different data mining tasks, allow preserving and presenting background knowledge. However, what seems to be lacking in most of the approaches, are the ways for taking the iterative and interactive nature of the data mining process into account (Mannila, 2000). Furthermore, none of the considered frameworks considers data mining in the context of an adaptive system that processes information.

In the next section we introduce an information systems development framework and then consider how data mining can be seen as an iterative and interactive development process within this framework.

REFERENCES


2. A Thesaurus Construction Method from Large ScaleWeb Dictionaries, Kotaro NAKAYAMA Takahiro HARA, Shojiro NISHIO, Dept. of Multimedia Eng., Graduate School of Information Science and Technology, Osaka University, 1-5 Yamadaoka, Suita, Osaka 565-0871, Japan {nakayama.kotaro, hara, nishio}@ist.osaka-u.ac.jp

3. Wikipedia Mining for An Association Web Thesaurus Construction, Kotaro Nakayama, Takahiro Hara and Shojiro Nishio, Dept. of Multimedia Eng., Graduate School of Information Science and Technology, Osaka University, 1-5 Yamadaoka, Suita, Osaka 565-0871, Japan. TEL: +81-6-6879-4513 FAX: +81-6-6879-4514 {nakayama.kotaro, hara, nishio}@ist.osaka-u.ac.jp

4. Web-Based Measure of Semantic Relatedness, Jorge Gracia and Eduardo Mena IIS Department, Univ. of Zaragoza, Mar’ia de Luna 1, 50018 Zaragoza, Spain {jogracia,emena}@unizar.es http://sid.cps.unizar.es

5. GENERALIZED MAXIMUM FLOW ALGORITHMS, Kevin Daniel Wayne, Ph.D., Cornell University 1999

6. The Google Similarity Distance, Rudi L. Cilibrasi and Paul M.B. Vit’anyi
