Weighting Relations in Social Networks Using the Web

Mizuki Oka Yutaka Matsuo

*1University of Tokyo

Measuring the weight of the relation between a pair of entities is necessary to use social networks for various purposes. Intuitively, a pair of entities has a stronger relation than another. It should therefore be weighted higher. We propose a method, using a Web search engine, to compute the weight of the relation existing between a pair of entities. Our method receives a pair of entities and various relations that exist between entities as input. It then outputs the weighted value for the pair of entities. The method explores how search engine results can be used as evidence for how strongly the two entities pertain to the relation.

1. Introduction

Social networks have garnered considerable interest because they offer the great potential for use in augmenting Semantic Web development. Various previous studies have addressed extraction of social networks from the Web [5, 7]. Most methods used in those studies are co-occurrence-based metrics used to compute the weight of the extracted relations among entities. Mika developed a system for extraction, aggregation, and visualization of online social networks for a Semantic Web community, called Flink [7], where the social networks are represented as a graph with a node representing a person and an edge indicating that a relation exists between the nodes. In that system, the relation weight is determined using the number of hits returned by a search engine, where the strength of the relation is measured through metrics such as the Jaccard coefficient. Using a similar approach, Matsuo et al. developed a system called Polyphonet [6]; in line with those studies, numerous studies have explored relation extraction to construct large-scale social networks from the Web [4, 1, 3, 2].

Given such extracted large-scale relational data, the weight of relations plays an important role in using the essential parts of social networks selectively for various applications. Such applications include searching for a valuable path from one person to another using the highly weighted relations and displaying the information (e.g., related entities) of a particular person in descending order of the weighted values. For example, if an entity *Steve Jobs* has a *CEO* relation with companies such as *Apple Computer Inc.*, *Pixar*, and *NeXT*, then the knowledge that the entity *Steve Jobs* has a prominent relation with the entity *Apple Computer Inc.* enables someone to use the extracted relational data selectively.

Although many studies have been proposed to extract large-scale relational data as described above, few studies have examined how to weigh each relation among entities beyond the co-occurrence based metrics [6]. Co-occurrence based metrics are based on the simple assumption that the co-occurrence of a pair of entities indicates the strength of their relation. Although this method provides a simple means to weigh relations, it neither takes the context differences into account nor offers clues about the strength of relations [8]. For example, it does not distinguish between (1) a pair of entities (e.g. people) that happen to co-occur because they participated in the same conferences but never co-authored a paper together (consequently, they have a *co-participant* relation), and (2) a pair of people who co-occur as a result of actually co-authoring papers (therefore, they have a *co-author* relation). Intuitively, a pair of people who share a co-author relation should be weighted more highly than others.

query	**target entity**	**result**
Apple Computer Inc., *CEO*	Steve Jobs	○
Yahoo!, *CEO*	Jerry Yang	×
Apple Computer Inc., *entrepreneur*	Steve Jobs	×
Yahoo!, *entrepreneur*	Jerry Yang	○
Steve Jobs *CEO*	Apple Computer Inc.	○
Jerry Yang *CEO*	Yahoo!	○
Steve Jobs *entrepreneur*	Apple Computer Inc.	○
Jerry Yang *entrepreneur*	Yahoo!	○

Table 1: Appearance of the target entity on the top ranked search result obtained using various queries.

Given this observation, we propose a novel method to weigh relations among pairs of entities. Given a pair of entities (A, B) and a keyword k, our method assigns a weight to a pair of entities (A, B) by analyzing the top N search results of a query composed of A and k. If the top N search results contain B, then the method regards it as evidence of people's common-sense knowledge that entities A and B share a relation of the keyword k. It then assigns a weight to the pair of entities according to the generality of the keyword k, which is measured according to its web hit counts. This method of weighting the relation among entities provides a basis for the weight of relations through the keyword, overcoming the shortcomings of the co-occurrence based metrics simultaneously: the unrelated entities such as a pair of people co-occurring a number of time on conference pages without having any relation are distinguishable from the pair of people who co-author several papers.

Contact: mizuki@cks.u-tokyo.ac.jp, matsuo@biz-model.t.u-tokyo.ac.jp

2. Method

2.1 Concept

As an exemplary scenario for our approach, we use two sets of relations (*Apple Computer Inc.*, *Steve Jobs*) and (*Yahoo!*, *Jerry Yang*) and a set of keywords *CEO* and *entrepreneur* that represent the relations. Given such data, our present goal is to weigh each pair of entities.

Table 1 shows results of an analysis of whether the page contains the target entity (e.g. *Steve Jobs*) in the top ranked search result of a query composed of the other entity (e.g. *Apple Computer Inc.*) and a keyword (e.g. *CEO*). For example, when the query "*Apple Computer Inc. CEO*" is issued, the top ranked search result contains the entity *Steve Jobs* (marked as ∘ in the table), although the entity *Jerry Yang* does not appear in the top ranked page (marked as × in the table) of the query "*Yahoo! CEO*". In contrast, when the keyword *entrepreneur* is used, the opposite result is obtained. The keyword *CEO* gives a much larger value than the keyword *entrepreneur* if one looks at the hit count of each keyword using the search engine. Based on the hypothesis described in Sect. 1, the method therefore assigns a higher weight to the relation (*Apple Computer Inc.*, *Steve Jobs*) than the relation (*Yahoo!*, *Jerry Yang*). The method also tests the other direction of the pair with, for example, the target entity of *Apple Computer Inc.* and the query composed of *Steve Jobs* and *CEO*. As presented in Table 1, all the combinations caused containment of the target entity in the top ranked pages; thereby, the weight is treated as equal. In the following section, we explain the precise steps of our proposed method.

2.2 Procedure

Our method for relation weighting of pairs of entities includes the following steps.

1 Collect a set of entity pairs.

2 Collect candidate keywords that describe the relations of pairs of entities.

3 Throw queries to a web-based search engine (e.g. Google) and examine the search results.

4 Calculate the weight of each relation accordingly.

Our method requires a list of entities (e.g., personal name, company name) and the lists of keywords that describe the relations among these pairs of entities as the input; it then outputs the weighted list of entity pairs.

3. Conclusion

Studies of relation extraction and keyword extraction have been conducted actively using the Web. Taking advantage of such recent studies and extending them, we proposed a method that weighs pairs of entities using the Web search engine.

Given a pair of entities and a set of keywords that are applicable among the entities, the method assigns a weight to a pair of entities (A, B) by analyzing the top N search results of a query composed of A and k. If the top N search results contain B, then the method regards it as evidence of people's common-sense knowledge that the entities A and B have a relation to the keyword k. It then assigns a weight to the pair of entities according to the generality of the keyword k, which is evaluated according to its web hit count.

Preliminary results on experiments which are not described in this paper due to space limitations demonstrate that the proposed metrics on weighting relations show positive correlations with the baseline rankings widely adopted by people. The results thus supports our hypothesis that if a pair of entities has a strong relation, then a general keyword can associate the two entities, and that the top search results of the Web search engines offers useful metrics to evaluate the relation.

Future studies will be undertaken to explore the possibilities of extending the proposed method to other types of relations, especially those that are related to people who would be useful in social networks. Such relations include basic keywords such as place, education, and office of employment to other entities such as hobbies and favorite products. We believe that weighting relations of various types will contribute to several social network applications such as finding a path from one person to another in terms of strong shared hobbies. Furthermore, seeking a seamless means to incorporate our method into the existing relation extraction and attribute extraction methods is an important subject for future work.

References

[1] L. Adamic and E. Adar. Friends and neighbors on the web. *Social Networks*, 25(3):211–230, 2003.

[2] A. Culotta, R. Bekkerman, and A. McCallum. Extracting social networks and contact information from email and the web. In *Proc. of the Conference on Email and Spam*, 2004.

[3] M. Harada, S. ya Sato, and K. Kazama. Finding authoritative people from the web. In *JCDL '04: Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries*, pages 306–313, New York, NY, USA, 2004. ACM.

[4] H. A. Kautz, B. Selman, and M. A. Shah. The hidden web. *AI Magazine*, 18(2):27–36, 1997.

[5] Y. Matsuo, M. Hamasaki, Y. Nakamura, T. Nishimura, K. Hasida, H. Takeda, J. Mori, D. Bollegala, and M. Ishizuka. Spinning Multiple Social Networks for Semantic Web. In *AAAI 06*, pages 1381–1387, 2006.

[6] Y. Matsuo, J. Mori, M. Hamasaki, K. Ishida, T. Nishimura, H. Takeda, K. Hasida, and M. Ishizuka. POLYPHONET: An Advanced Social Network Extraction System from the Web. In *WWW '06*, pages 397–406. ACM Press, 2006.

[7] P. Mika. Flink: Semantic Web technology for the extraction and analysis of social networks. *Web Semantics: Science, Services and Agents on the World Wide Web*, 3(2-3):211–223, 2005.

[8] J. Mori, M. Ishizuka, and Y. Matsuo. Extracting keyphrases to represent relations in social networks from web. In *IJCAI 07: International Joint Conference on Artificial Intelligence*, pages 2820–2827, 2007.