Chapter 3
Semantic and Social Spaces: Identifying Keyword Similarity with Relations

Yun Huang, Cindy Weng, Baozhen Lee and Noshir Contractor

Introduction

With the development of Web 2.0 technologies, semantic contents and social interactions have become tightly integrated in online social networks and social media such as blogs, micro blogs, and social tagging. Moreover, user generated content on Wikipedia and citizen science sites (e.g., Scitable at Nature Publishing Group), has begun to organize and even generate knowledge from crowd participations.

Most content generated by users is noisy, ambiguous, and unstructured because of the voluntary nature of contributors and varying reliability of information resources. On the other hand, complex human interactions can provide the rich information to reveal the expertise and credibility of users and in turn optimize the process of information retrieval on the Web. A key problem in constructing and mining semantic spaces is how to utilize users' preference information in the process of extracting information structures from unstructured data sources and construct a relevant concept similarity network (Steyvers and Tenenbaum 2005). By jointly considering text and relational data, we propose to analyze multiple dimensions of human expertise and behavior.

This chapter proposes a three-layer framework to integrate semantic and social networks and to reveal people's expertise based on their words and relations. To demonstrate the value of user preference in semantic analysis, we use social tagging activities on CiteUlike as an example to illustrate the potential of utilizing social relations on identifying similar concepts.

Y. Huang (✉) · C. Weng · N. Contractor
The Science of Networks in Communities (SONIC) research group,
Northwestern University, Evanston, IL, USA
e-mail: yun@northwestern.edu

B. Lee
School of Economics and Management, Jiangsu University of Science
and Technology, Zhenjiang, China

E. Bertino, S. A. Matei (eds.), *Roles, Trust, and Reputation in Social Media Knowledge Markets,* Computational Social Sciences, DOI 10.1007/978-3-319-05467-4_3,

Semantics Meets Social Networks

Semantic networks represent semantic relations among concepts using linguistic information, such as documents and keywords. There are a few approaches in constructing concept similarity networks. Formal concept analysis (FCA) is a principled way to derive formal ontology from a collection of objects and their properties (Ganter et al. 2005). The binary relations between objects and attributes reveal formal concepts and concept hierarchy. Similarly, latent semantic analysis (LSA) applies the singular value decomposition (SVD) method to reduce dimensions and construct concept networks based on the frequency relations between documents and keywords (Deerwester et al. 1990). Centering resonance analysis (CRA) (Corman et al. 2002) uses keyword adjacency relations in sentences to understand how the keywords are being used in a specific document. Whereas keyword frequency methods create insights based on a "pile of words," CRA mainly adopts the betweenness concept of social networks. These approaches mostly focus on content of documents and neglect the socio-technical context such as user preference and why and how frequently people use the keywords.

To reflect the influence of user preference information on the process of constructing concept similarity networks, tripartite network models such as high-order singular value decomposition (HOSVD) (Omberg et al. 2007) utilize three types of elements, e.g. authors, keywords, and documents. While incorporating an additional dimension with content information, the users are treated as an additional type of nodes providing more association information but not as a human agents whose preferences, expertise, and relations change the use of keywords and documents.

There are some recent technical approaches to integrate semantic and social networks. The semantic social network (SSN) (Downes 2004) has extended the ontology of Semantic Web to online social networks. Using the Resource Description Framework (RDF), a conceptual description of information in triples and XML, every type of entity and relation is defined by descriptive vocabularies and the ontology provides a new view of Web information space connecting people and resources. As a formal method, SSN utilizes reasoning and deduction to retrieve information among people with related interests. In a more generic setting, heterogeneous information network (Sun et al. 2009) considered the collection of social and semantic networks as a hybrid network of multiple types of entities and multiple relations. Mining a particular meta path, i.e. a sequence of relations among different entity types, reveals the potential similarity structures in a complex network. Both approaches take a symmetric and abstract view of entity types. For example, a user is not different from a keyword unless they are specifically characterized by ontology or meta-paths. Without a conceptual framework, these approaches are limited in their ability to establish a systematic way to combine social and semantic networks.

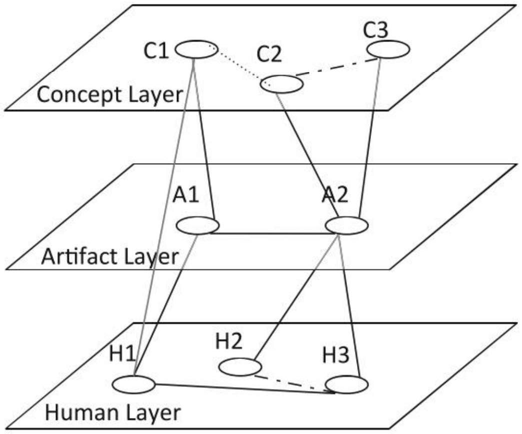

Fig. 3.1 Three-layer multi-dimensional framework

Three-Layer Framework for Multidimensional Networks

In order to reveal the inherent influences in large complex networks, we classify entities in a multi-dimensional network using three categories: human, artifacts, and concepts; and thus construct a three-layer framework representing heterogeneous knowledge and social networks (Contractor et al. 2011). Figure 3.1 describes the three layers and the relations within and between layers.

The *concept layer* represents the content domain of a knowledge and semantic network and consists of all knowledge entities including keywords/tags, properties, classes, topics. The links between entities are the logic relations defined by their ontologies. The semantic networks are either directed (e.g. semantic trees characterizing the concept hierarchy) or non-directed (e.g. concept similarity/sibling networks).

The human layer represents social networks and consists of human agents who can make decisions and actions. Each agent has a certain profile (for example status, preference, and expertise) which potentially affects its behavior. Agents could be individuals or aggregates of individuals such as groups, organizations, countries, etc. The links between two agents are their social relations and interactions such as friendship and communication. The structural tendencies of these social relations reflect the underlying motivations for creating and maintaining links such as homophily and proximity (Monge and Contractor 2003).

The artifact layer represents a collection of physical and information artifacts created by human agents—web pages, articles, products, and events. Artifacts are linked by various connections based on the content or usage such as web page hyperlinks, article citations, and product promotion events. Artifacts also act as intermediaries connect concepts with human agents. Some artifacts are associated with concepts (e.g. document-keywords and product-properties), while others link to human activities and transactions (e.g. user's web page access and product purchase).

The association relations and transactions can be projected to the concept and human layers and be used to generate derived relations. The combination of different relations can produce more information about user behavior and the knowledge domain. For example, suppose person H1 is an editor of a journal (A1) that focuses on “social network analysis” (C1). Therefore the information entity A1 establishes a relation between H1 and C1. Suppose A2 is a research paper, with keywords “friendship” (C2) and “recommender systems” (C3) in the journal A1 coauthored by H2 and H3. The artifact A2 generates many derived relations such as co-authorship between H2 and H3, keyword co-occurrence between C2 and C3, H2 and H3’s expertise indicated by keywords C2 and C3, potential interests of keyword C2 for the journal A1, etc.

The three-layer framework is very flexible in preserving various types of relations: semantic networks in the concept layer, social networks in the human layer, and association and transaction relations in the artifact layer and between layers. The concept and artifact layers represent the application scenario of content and semantic analysis, and the human and artifact layers represent the scenario of social interactions and sociomateriality (Contractor et al. 2011). Based on this framework, a multi-dimensional network can be represented as a tensor (e.g. a 3-dimensional array) and simplified through dimensionality reduction for higher-order factor analysis. For example, the higher-order generalization of SVD (HOSVD) for tensors (Lathauwer et al. 2000) is used efficiently in independent component analysis (ICA) and converts a given N-dimensional tensor into a full orthonormal system in a special ordering of singular values. It is capable of extracting clear and unique structures underlying the given multi-dimensional network (Lu et al. 2011).

Contributions of Social Relations in Identifying Concept Similarity

To demonstrate the utility of incorporating information about social relations to identify similar topic words, we use a sample data set in the social tagging website CiteULike. We compare different LSA methods. In CiteULike’s interest group “Blog and Wiki Research,” there were 2961 tagged documents between November 2004 and August 2010. From these documents, we selected 69 research papers that had been tagged by at least ten users as the test case. These documents were tagged with 169 different tags by 145 users.

We evaluated the performance of three approaches, SVD, HOSVD, user-oriented SVD (UoSVD), to demonstrate the contribution of information at different layers in the three layered framework in semantic analysis.

Latent semantic analysis (LSA) takes a sample of documents as a term-by-document matrix where each cell indicates the frequency with which each term (rows) occurs in each document (columns). Using SVD, the matrix is reduced into a low dimensional vector space in which each term and document is identified by a vector. Thus, the distance between a pair of term vectors provides a similarity measure between the two terms. In the case of social tagging, users choose some tags to annotate relevant information resources. Similarly higher-order SVD (HOSVD) uses a 3-dimensional array (term-by-document-by-user) to include user information with terms and documents and the relations among the three types of entities are used to construct concept similarity in a reduced vector space.

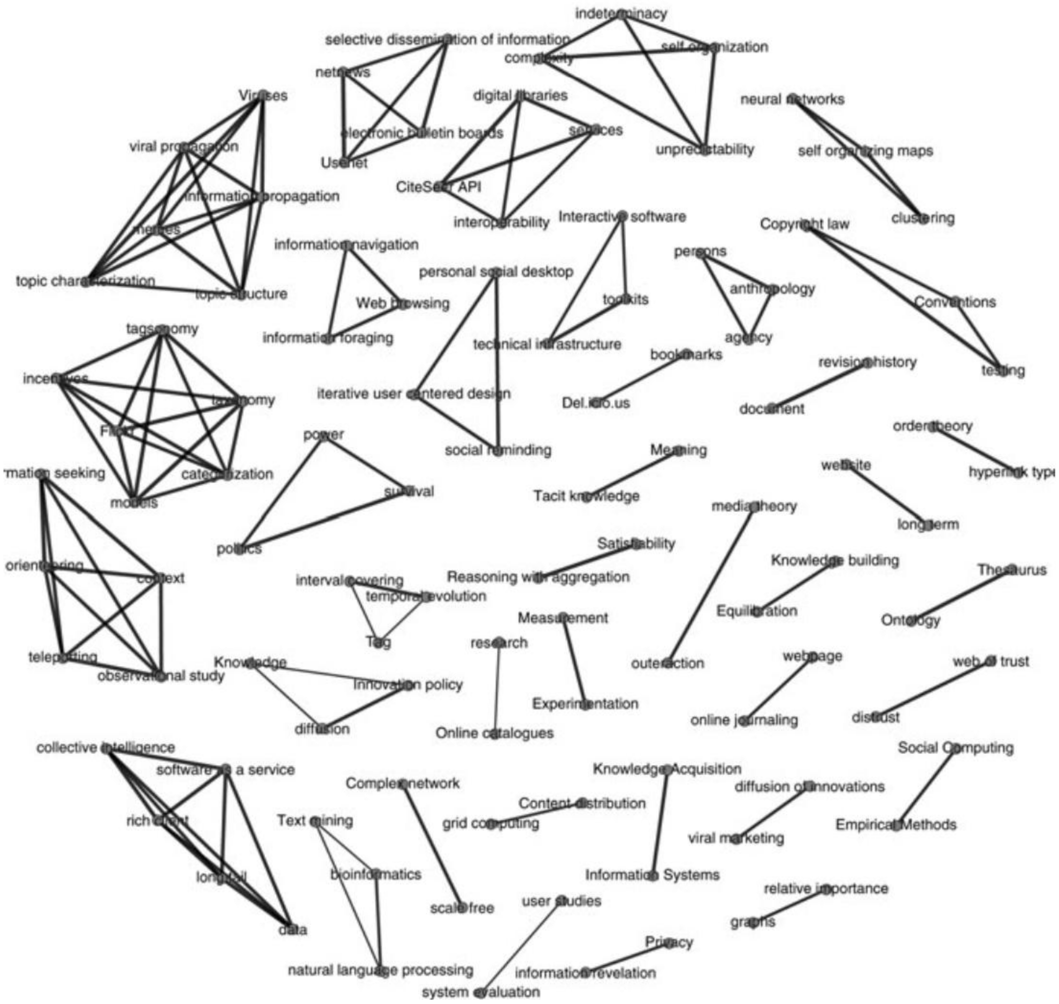

Fig. 3.2 Concept similarity network of sample data based on UoSVD (similarity > 0.4, and isolates removed)

Since the concepts depend on not only their relations to relevant documents but also the types of users who use it, the similarity between tags should consider preference information of corresponding taggers. Therefore we construct user-oriented information explicitly through a term-user matrix in which each cell is a term frequency–inverse document frequency (tf-idf) measuring the relative frequency a tagger used in the collection of the documents. In this user-oriented approach, we use both term-document and term-user matrices via traditional SVD to estimate similarity of the concepts.

Using the sample dataset, the three approaches generated keyword similarity networks with similar network structures. Figure 3.2 visualizes the concept similarity network produced by the UoSVD model. Only the links with similarity values larger than 0.4 are included and the graph shows clusters of tags with their similarity indicated by link width. No isolated tags are included.

In actuality, the link weights in the concept similarity network should be bigger between tags conceptually similar, and smaller among different tags. Therefore, the evaluation criteria should consider not only the consistency of concept similarity in

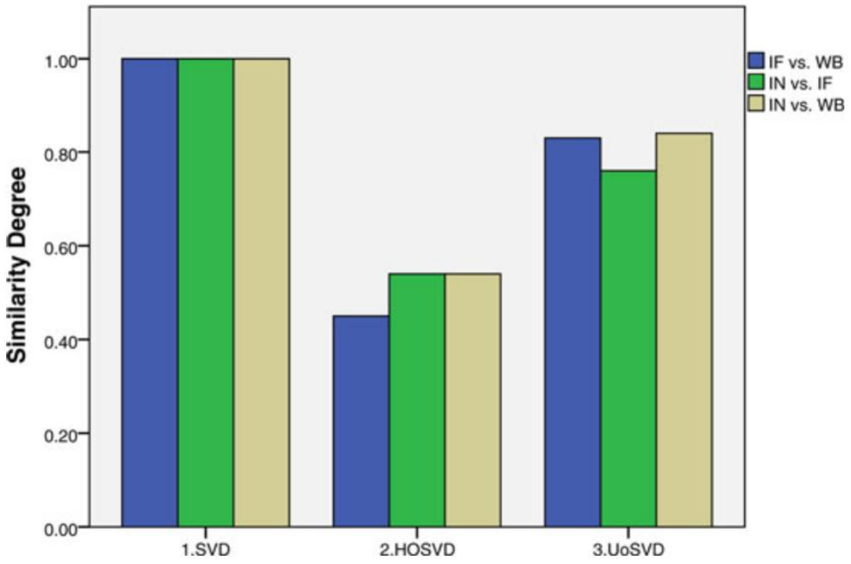

Fig. 3.3 Similarity scores among three related tags

this sample set extracted from one interest group, but also consider the discrimination power to separate different concept clusters in the set.

Figure 3.3 shows the similarity scores between each pair of elements in the cluster—information navigation (IN), information foraging (IF), web browsing (WB)—as an example. The similarity scores among the three tags based on the UoSVD have a larger power to discriminate the tags from each other, whereas the traditional SVD model produce almost the same similarity scores for each pair of tags. The UoSVD model that considers the influences of user preference can be used to discriminate the relations among relevant concepts more accurately in intra-cluster.

As a comparison, Fig. 3.4 shows similarity scores between each element in the cluster of the three tags and an unrelated term "peer to peer" (P2P). Again, we find that the UoSVD is more effective to detect the differences among concepts in different clusters than the traditional SVD and HOSVD models. Furthermore, the UoSVD has a much higher signal ratio of similarity scores of related tags vs. similarity scores of irrelevant tags and therefore has a much better performance to identify true associations among tags. On the other hand, using the same amount of information, the HOSVD achieves the worst results in discriminating tags. This suggests that user behavior and social relations provide a different mechanism in influencing semantic networks, and user information does not reveal the association of tags directly.

Based on accuracy evaluation methods in the literature (Breese et al. 1998), we evaluate the consistency of concept similarity in the concept similarity network using the variance scoring metric. The expected variance of concept similarity between tag i and other tags is defined as

$$V_i^2 = \frac{1}{m-1} \sum_{j}^{m-1} (S_{ij} - \frac{1}{m-1} \sum_{j=1}^{m-1} S_{ij})$$

where S_{ij} is the similarity score between tags i and j. The overall consistency of all tags equals to one minus the average variance of all tags.

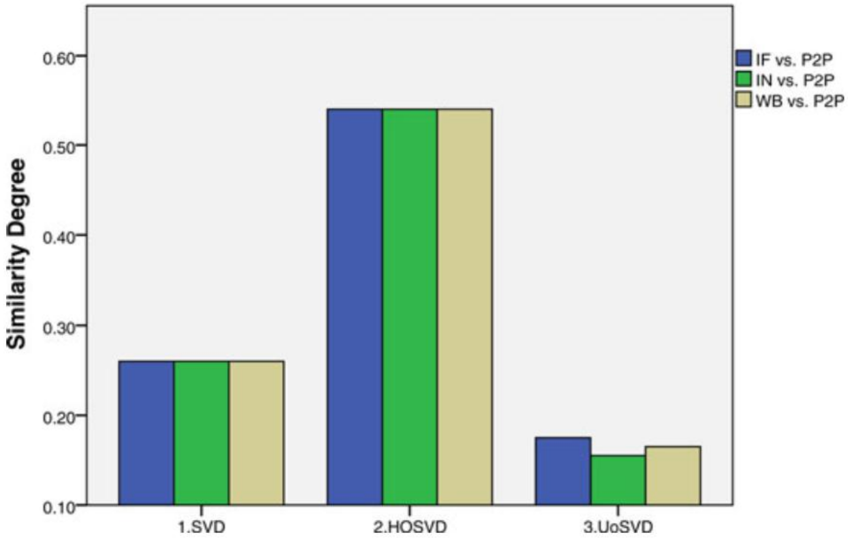

Fig. 3.4 Similarity scores between each of the three tags and an irrelevant tag "peer to peer (*P2P*)"

We evaluate the discrimination of concept similarity in the concept network using the range scoring metric. The expected range of similarity for tag *i* is defined as

$$R_i = \frac{S_{ij}^{\max} - S_{ij}^{\min}}{S_{ij}^{\max}} \forall j$$

where S_{ij} is the similarity score between tags *i* and *j*, and $S_{ij}^{\max}$ is the maximum possible similarity score between tag i and other tags, and $S_{ij}^{\min}$is the minimum similarity degree.

We consider the whole dataset as one cluster and calculate the average variance of similarity for each tag in concept similarity networks generated by different algorithms. Figure 3.5 illustrates the overall consistency and average discrimination scores of concept similarity in the three approaches. The results show that the UoSVD, which utilizes user preference information, produces more consistent similarity scores and has a slightly better discrimination power compared to the SVD model. The size of the differences is very small because we calculated the overall consistency and discrimination scores using all possible pairs of tags. Since many tags are not related and most similarity degrees between tags are zero, the two evaluation scores are highly inflated. More detailed comparisons among different clusters, like the three tag case we illustrated above, would show further significant differences.

Conclusion and Future Work

This chapter discusses the importance of integrating social relations and semantic networks in the discovery of topics and similar concepts in social media. Instead of using a generic network model and treating all types of nodes and relations equally in a heterogeneous network, we propose a three-layer model to characterize the uniqueness of semantic and relational information as well as their interconnections. Using

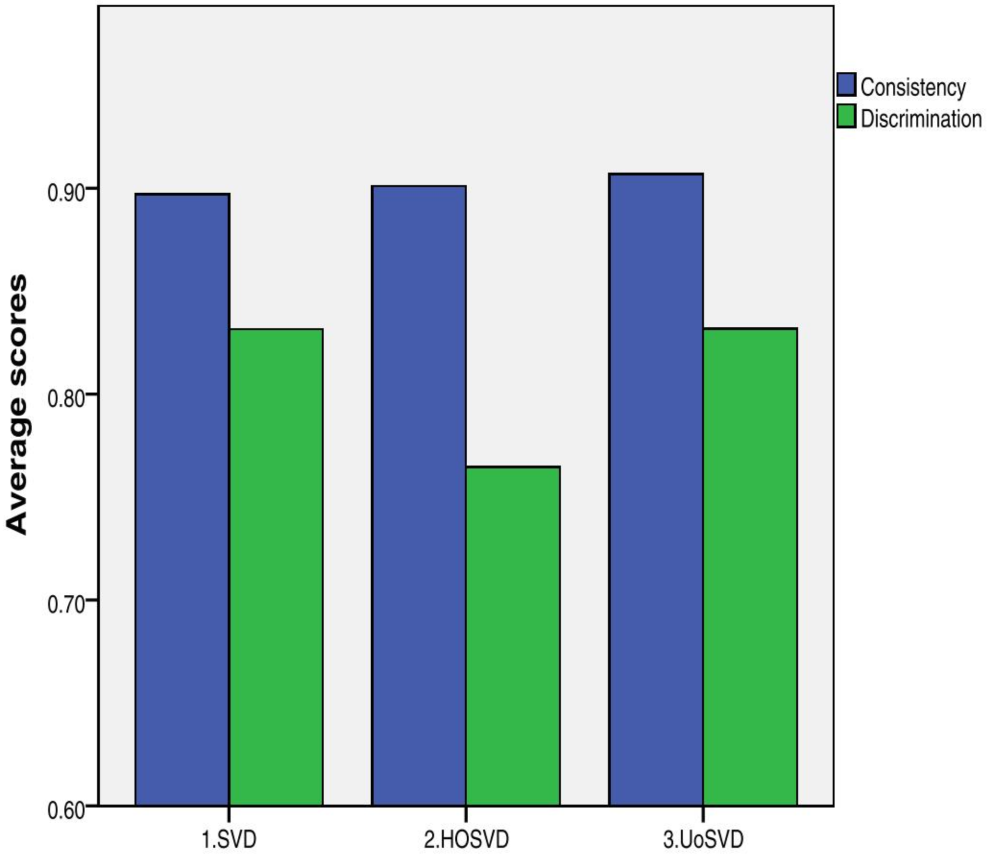

Fig. 3.5 Overall consistency and average discrimination scores of concept similarity in the three approaches

a sample case with social tagging in CiteULike, we demonstrate that the correct use of human preference and relational information can help identify similar concepts and topics. User-oriented information captures people's expertise, motivation, and preference in participating information consumption and production online, and therefore, it potentially affects the structure of the semantic networks.

The three-layer framework we introduce captures the essential structure of many application scenarios, such as scientific publication with authors, documents and keywords, Wikipedia with contributors, wiki articles and concept items, and marketing with consumers, products and features. The increasing practice of social networking makes the semantic space of online topics and content more dynamic. Further advanced methods to discover the latent social space underneath social relation networks have become a key challenge to further integrate the social and semantic networks.

Acknowledgements This study is supported by National Science Foundation (No. CNS-1010904, OCI-0904356, & IIS-0841583), Army Research Lab (W911NF-09-02-0053), the National Natural Science Foundation of China (No. 71273121/G031401).

References

Breese, J. S., Heckerman, D., & Kadie, C. (1998). *Empirical analysis of predictive algorithms for collaborative filtering*. Paper presented at the the fourteenth annual conference on uncertainty in artificial intelligence.

Contractor, N., Monge, P., & Leonardi, P. (2011). Multidimensional networks and the dynamics of sociomateriality: Bringing technology inside the network. *International Journal of Communication, 5,* 682–720.

Corman, S. R., Kuhn, T., McPhee, R. D., & Dooley, K. J. (2002). Studying complex discursive systems: Centering resonance analysis of communication. *Human Communication Research, 28*(2), 157–206.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science and Technology, 41,* 391–407.

Downes, S. (2004). The semantic social network. http://www.downes.ca/post/46. Accessed 24 Oct 2013.

Ganter, B., Stumme, G., & Wille, R. (Eds.). (2005). *Formal concept analysis: Foundations and applications*. Lecture notes in artificial intelligence (Vol. 3626). Berlin: Springer.

Lathauwer, L. D., Moor, B. D., & Vandewalle, J. (2000). A multilinear singular value decomposition. *SIAM Journal on Matrix Analysis and Applications, 21*(4), 1253–1278.

Lu, H., Plataniotis, K. N., & Venetsanopoulos, A. N. (2011). A survey of multilinear subspace learning for tensor data. *Pattern Recognition, 44*(7), 1540–1551.

Monge, P., & Contractor, N. (2003). *Theories of communication networks*. Oxford: Oxford University Press.

Omberg, L., Golub, G. H., & Alter, O. (2007). A tensor higher-order singular value decomposition for integrative analysis of DNA microarray data from different studies. *Proceedings of the National Academy of Sciences, 104*(47), 18371–18376.

Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science, 29*(1), 41–78.

Sun, Y., Yu, Y., & Han, J. (2009). *Ranking-based clustering of heterogeneous information networks with star network schema*. Paper presented at the ACM SIGKDD international conference on Knowledge Discovery and Data Mining (KDD'09).