Similarity Measure for Social Networks – A Brief Survey

Ahmad Rawashdeh and Anca L. Ralescu
EECS Department, ML 0030
University of Cincinnati
Cincinnati OH 45221-0030, USA
rawashay@mail.uc.edu, Anca.Ralescu@uc.edu

Abstract

Social networks play an increasing role in many areas of computer science applications. An important aspect of these applications relies on similarity measures between nodes in the network. Several similarity measures, described in the literature are surveyed here with the goal of providing a guide to their selection in various applications.

Introduction

Social networks represent a particular domain as a collection of nodes/profiles and links between them. Common operations in social networks, such as link prediction, community formation, browing, are driven by a similarity measure between nodes. Node similarity can be viewed as similarity between strings, whose definition/ evaluation can be traced to work on information retrieval (Findler and Van Leeuwen 1979).

Often similarity measures are defined as decreasing functions of a distance metric. For example, two of the string metrics used most often are *editDistance* (Lin 1998) and *trigrams* (Bahl, Jelinek, and Mercer 1983). For finite strings x and y the edit distance is defined as

$$d_{edit}(x,y) = \min\{\gamma(S) | S \text{ is a en edit sequence taking } x \text{ to } y\} \quad (1)$$

where γ denotes the cost of an edit operation (deletion, insertion, replacement), and for the sequence of edit operations $S = \{s_1, \ldots, s_n\}$, $\gamma(S) = \sum_{i=1}^{n} \gamma(s_i)$. The trigram distance for two sequences x and y is defined as:

$$d_{tri}(x,y) = \frac{|tri(x) \cap tri(y)|}{|tri(x) \cup tri(y)|} \quad (2)$$

where $tri(x)$ denotes the collection of trigrams (ordered substrings of length 3) of x, and $|tri(x)|$ denotes the number of trigrams of x. Then the similarity measures corresponding to (1) and (2) are defined as in equations (3) respectively (Lin 1998).

$$sim_a(x,y) = \frac{1}{1 + d_a(x,y)} \quad (3)$$

where $a \in \{edit, tri\}$.

Given a profile of a network node, finding similar profiles has been investigated by many researchers (Yang et al. 2012), (Huang and Lai 2006), (Pan et al. 2010), (Symeonidis, Tiakas, and Manolopoulos 2010). Automating this task may help when browsing large collections of data: instead of searching through a large network to find candidate profiles, a *similarity aware browser* can suggest them by considering similarities along some features. Applications of such browsers include social networks (e.g., Facebook, LinkedIn), as well as other networks (e.g., recommending systems)

For many people, day-to-day interaction has been replaced by instant messages, likes (or favorite), and share (retweet) through social networking websites such as Facebook, MySpace, Twitter, YouTube, and Orkut [1]. In particular, by the end of 2010, Facebook had in excess of 1.2 billion users (Facebook 2010). Many people have turned to such websites to communicate with friends or make new connections. This increase in internet usage has raised many questions concerning the privacy of these users, since they upload their personal media content (photos and videos) and they share their personal opinions on various topics (Díaz and Ralescu 2012).

The motives for participating in online social networks could be understood from the study of psychology. See for example (Heidemann, Klier, and Probst 2012) where the definition of social networks, their characteristics, as well as what motivates participation in them is presented.

Formally, a social network can be represented as a graph (Díaz and Ralescu 2012), that is a collection of *nodes*, or *profiles*. Similarity between nodes could be based on node attributes(textual) and/or edges/links(structure).

Some similarity measures consider the common neighbors of nodes (Jeh and Widom 2002), while others allow nodes to be similar even when they do not have common neighbors (Leicht, Holme, and Newman 2006). Some similarity measures only consider the link similarity of length two, others define similarity based on longer paths (Leicht, Holme, and Newman 2006), while others are defined as *the number of paths of varying length* between them (Papadimitriou, Symeonidis, and Manolopoulos 2012). Applications of node similarity are different and they have inspired researchers to explore different approaches for evaluating it.

[1] www.facebook.com, www.myspace.com, www.twitter.com, www.youtube.com, www.orkut.com

For example, some work combines similarity from Wordnet with a *vector cosine similarity* (Rawashdeh et al. 2014) to find similarity of profiles in Facebook.

Several similarity measures have been introduced including, Jaccard (biology) (Jaccard 1912), cosine, min (Leicht, Holme, and Newman 2006), Sorensen, Adamic Adar (Adamic and Adar 2003), and resource allocation (Zhang et al. 2010). Also, *PageSim*, a method to measure the similarity between web documents was proposed in (Lin, King, and Lyu 2006), based on *PageRank* score propagation. *PageSim* was evaluated against standard information retrieval similarities TF/IDF, which were considered to be the ground truth. Most of the similarity measures described in the literature are knowledge dependent. However, the authors in (Lin 1998) describe an independent definition of similarity in terms of information theory. A list of similarity properties (axioms) was included in (Burkhard and Richter 2001).

Semantic Similarity

Research in finding the *semantic similarity* between concepts using knowledge such as Wordnet or between words in the semantic web has been reported in several papers (Li, Bandar, and McLean 2003), (Ilakiya, Sumathi, and Karthik 2012). Semantic similarity measures have been classified into (1) feature based, (2) information content (which relies on counting the number of occurrences of a word in corpora for instance), (3) hybrid, and (4) path/ontology measures (which counts the number of edges/nodes between two concepts) (Elavarasi and Menaga 2014).

The path similarity measure is based on the structure of the taxonomy of the conceptual relationships (ontology hierarchy) and it is sensitive to the quality of the taxonomy of concepts. This determines how the semantic similarity measure is quantified. Edge counting methods suffer from irregularities in path lengths between different concepts so one must proceed with caution when using them.

Approaches based on information content combine corpus statistics and taxonomy structure (Jiang and Conrath 1997). The results report that the information content measures perform better than edge only based measures. It is worth noting that most studies that use Wordnet only consider the *is-a* relationship (hyponymy/hypernymy) (Li, Yang, and Park 2012).

A comparison between the three different similarity measures was discussed in the paper (Pirró 2009). The authors have pointed out that approaches that rely on statistics of word occurrences, within the corpora, require intensive computations, and thus are not practical when the corpora is large or is different from the one used to find information content.

Wordnet is a free lexical database that organizes English words into concepts and relations between them. English nouns, verbs, adjectives, and adverbs form hierarchies of *synsets* with relations connecting them. A *synset* is the hierarchy determined by the *hypernym* (is-a) relationship.

Wordnet-Similarity is a Perl package for calculating the similarity between concepts using Wordnet (Pedersen, Patwardhan, and Michelizzi 2004).The package implements six different similarity measures, three of which are based on information content and the remaining three are edge based similarity measures.

The work described in (Mabotuwana, Lee, and Cohen-Solal 2013) uses cosine similarity in conjunction with the SNOMED CT ontology to evaluate similarity between words. Similarly, in (12) cosine similarity is also used, however, in conjunction with Wordnet which is a more general ontology than the SNOMED CT ontology. In addition, the approach described in (12) finds the similarity between sentences not just words. Therefore tools of natural language processing(NLP) are considered.

Problem description and evaluation metrics

The problem of finding node similarity can be concisely stated as follows: given a node with attributes and possibly a set of structural attributes represented as connections find the set of nodes which are similar to it. Using the formalism of graph theory, a social network is defined as a graph $G = (V, E)$, where V, the set of vertices represents nodes in the the network, and E the set of edges, represents the links in the network. Thus the similarity problem is to find all pairs of similar vertices (v_i, v_j), $v_i, v_j \in V$, based either on the node profiles (node attributes) or the set of edges E.

Motivation for finding similarity

The problem of finding similar objects has its root in clustering, collaborative filtering, and search engines(Ganesan, Garcia-Molina, and Widom 2003). Finding similar objects can be used to predict links in data networks (Lü and Zhou 2011). There are two approaches for link prediction either local or global link structure (overall path). Also, finding similar objects may be used to recommend items for a customer or friends for a particular person based on commonality between the objects attributes (Yang et al. 2014). Prior to recommender systems, the problem of finding similar objects was also studied in information retrieval (Lin, King, and Lyu 2006), similarity is used to cluster documents (Zhou, Cheng, and Yu 2009). Measures of similarity used for this purpose include content-based, title-based, and keyword-based (Xiao 2012) measures. An example of using similarity for clustering is collaborative filtering (Jeh and Widom 2002). Zhou, Cheng, and Yu proposed an algorithm to clustering objects using attributes and structure where the attribute of a node and the structure are seemingly conflicting or at least independent (Zhou, Cheng, and Yu 2009). Different similarities measures have been used in biology, ethnology, taxonomy, image retrieval, geology, and chemistry (Choi, Cha, and Tappert 2010), as well as in the biomedical field (Mabotuwana, Lee, and Cohen-Solal 2013). Applications of finding similarity in data include (Li et al. 2010) neighborhood search, centrality analysis, link prediction, graph clustering, multimedia captioning, related pages suggestion in search engines, identifying web communities, friends suggestion in friendship network (Facebook or MySpace), movies suggestion, item recommendation in retail service, scientific and web domains in general.

Table 1: Two Facebook profiles

Profile ID	Profile	Friends IDs
xxxxxxx773.html	Comedy, Action films American, EL EL	1, 2, 3, 10
xxxxxx432.html	Haunted 3D, Saw, Transformers, Pirates of the Caribbean, Mind Hunter	3, 4, 5, 9, 10

Node and other similarity measurements

When the graph structure is considered, the similarity is based on node and edge properties. When general ontologies or domain knowledge are used, then *semantic similarity measures* are used. Furthermore, depending on the context, similarity between words, documents, or between profiles (nodes) are used (Symeonidis, Tiakas, and Manolopoulos 2010), (Naderi and Rumpler 2007). According to their types, similarity measures in networks can be classified as:

- **Structural similarity (link-based).** In this type of similarity, the links between the nodes in the graph are examined; the links can represent: co-authorship, friendship, payment, etc. It has been shown that when compared with respect to the human judgment they are better than text similarities (content)(Li et al. 2010). An example of structural similarity, which takes into account the neighbors of the pair of vertices under consideration is defined in (Leicht, Holme, and Newman 2006). Table 4 shows a list of structural similarity measures.
- **Content similarity (text-based).** In this type of similarity, the attributes of the node in the graph are examined. Content similarity of a friendship website could possibly be based on birth date, hobbies, movies interest, and age. One way to capture content is by the use of user-defined tags (e.g., tags were considered to represent the content of a movie of interest to the user while building a group profile). Based on tag similarity, a recommendation algorithm can be developed (Pera and Ng 2013).
- **Keyword similarity (word-based)**. Like for tag similarity, node similarity may be defined based on the similarity between node representing collections of words: keywords. An example of keyword similarity is the *forest model* described in (Bhattacharyya, Garg, and Wu 2011), where the keywords were arranged in a hierarchical structure to form trees of different heights. Wordnet was then used to find the semantic relationship between the keywords.

Tables 2 and 3 show an example of two Facebook profiles and their similarities as evaluated by a group of six users. In table 1 for each profile ID, the movies interest, and the list of friends, are included. This data is a Facebook snaphot, where the friend IDs are synthetic data. The scores in table 3 range from $[-2, 2]$ with negative score indicating dissimilarity and positive scores indicating similarity.

Node similarity

The similarity measures compared in (12) are Wordnet-Cosine, Word Frequency Vector, Symantic Categories,

Table 2: Two Facebook profiles

Dataset	Facebook
Profile-1 ID	100000060663828.html
Movies Interest	Captain Jack Sparrow, Meet The Spartans, Ice Age Movie, Spider-Man
Profile-2 ID	100000067167795.html
Movies Interest	Clash of the Titans, Ratatouille, Independence Day, Mr. Nice Guy, The Lord of the Rings Trilogy (Official Page)

Table 3: Human judgement on the Facebook profiles shown in Table 2

Person	Similarity score in $[-2, 2]$
User 1	1
User 2	0
User 3	2
User 4	1
User 5	2
User 6	0.8
Average	**1.13**

and Set similarities. For the Wordnet-Cosine measure, a node profile X is represented by the vector $D_X = [D_{x1}, \ldots, D_{xn}]$, where D_{xi} denotes the distance, in the hierarchy of concepts, between the ith word in the user profile X and the top concept *entity*, obtained by using Wordnet. The Wordnet-Cosine similarity is then defined as shown in equation (4)

$$Sim_W(X, Y) = \cos(D_X, D_Y), \tag{4}$$

For the WFV similarity measure, a node profile X is represented by the vector $F_X = [F_{x1}, \ldots, F_{xn}]$, where F_{xi} denotes the denotes the frequency of the ith word in the dataset. The Word Frequency Vector similarity is then defined as shown in equation (5)

$$Sim_{WFV}(X, Y) = \cos(V_X, V_Y), \tag{5}$$

The Symantic Category similarity measure is defined as shown in equation (6).

$$Sim_{SC}(X, Y) = \cos(SC_X, SC_Y), \tag{6}$$

where

$$SC_X = [f_A(X) | A \in \{NN, NNS, NNP, NNPS\}],$$

and $f_A(X)$ denotes the frequency of A in X.

Finally, the Set similarity is defined as shown in equation (7).

$$Sim_S(X, Y) = \frac{|S_X \cap S_Y|}{|S_X \cup S_Y|}, \tag{7}$$

where $S_X = \{S_{xi} | i = 1, \ldots, n\}$ is the set of parents for the ith word in the user profile X obtained by using Wordnet.

Table 4: Node and Link similarities

Node Similarity			
Wordnet Cosine	Set	Semantic	Word Frequency Vector
0.862795963	0.0659340066	0.877526909	0.74900588
Link Similarity			
Slaton	Jaccard	Hub Promoted Index	Hub Depressed Index
0.423	0.285	0.5	0.4

Edge similarity

Several structural similarity measures, based on edges are shown in equations (8) - (11), where $\Gamma(X)$ denotes the set of neighbors of X, and K_X is the degree of node X:

$$Sim_{Salton}(X,Y) = \frac{|\Gamma(X) \cap \Gamma(Y)|}{\sqrt{K_X \times K_Y}} \tag{8}$$

$$Sim_{Jaccard}(X,Y) = \frac{|\Gamma(X) \cap \Gamma(Y)|}{|\Gamma(X) \cup \Gamma(Y)|} \tag{9}$$

$$Sim_{HPI}(X,Y) = \frac{|\Gamma(X) \cap \Gamma(Y)|}{\min\{K_X, K_Y\}} \tag{10}$$

$$Sim_{HDI}(X,Y) = \frac{|\Gamma(X) \cap \Gamma(Y)|}{\max\{K_X, K_Y\}} \tag{11}$$

Table 4 shows the similarities between the two Facebook profiles (shown in Table 1). The top portion of the table, shows the node similarities computed according to equations (4)-(7), while the bottom portion shows the edge-similarities computed according to equations (8)-(11). It can be seen from table 4 that, with the exception of set similarity, node/profile semantic similarity measures exceed the measures based on links. The maximum node similarity is attained by Wordnet Cosine, which exceeds 0.86, while the lowest similarities are attained by set similarity and Jaccard similarity respectively. Note that this is (or it should not be) surprising, for the Wordnet Cosine captures the similarity of meanings based on the Wordnet hierarchy. The Jaccard similarity is an index of intersection of the set neighbors, without any semantic analysis of their meaning. The spirit of set similarity is actually quite close to that of the Jaccard similarity, as it provides the index of intersection of node parents (which are, of course among the neighbors) of the nodes being compared. In a browsing application, setting a threshold, α, on the similarity of items returned given a query, if $\alpha \geq 0.5$ three of the node similarity measures will output the two profiles as similar. And in fact, the same result would hold when $\alpha = 0.74$. By contrast, with $\alpha = 0.5$ only one of the link similarity measures would output them as similar.

Global Structural Similarities

Structural similarity can be classified according to three perspectives: (i) local vs. global, (ii) parameter-free vs. parameter-dependent, and (iii) node-dependent vs. path-dependent (Lü and Zhou 2011). In general, global structural similarity measures, some of which are listed below, aim to evaluate the similarity between two nodes in the context of the whole network.

$SimRank$ is a general approach for finding similarity between objects, based on structural features (Jeh and Widom 2002). Two objects are considered to be similar if they are related to similar objects. The authors state that performance was out of scope in their experiment. $SimFusion$ (Xi et al. 2005) finds the similarity between two objects by considering evidence from multiple sources (data spaces). One of the differences between $SimRank$ and $SimFusion$ is that $SimFusion$ uses two *random walker* models while $SimRank$ uses a *random Surfer-Pairs* model (Xi et al. 2005). A non-iterative version of $SimRank$, was shown to have improved performance (Li et al. 2010) .

$P-Rank$ (Zhao, Han, and Sun 2009) extends $SimRank$ by taking into consideration the in-links and out-links relationship when calculating the similarity. According to $P-Rank$, which expands the definition of $SimRank$, "two entities a and b are similar, if they are referenced by similar entities"and "if they also reference similar entities". $E-rank$ of two nodes, measures probability of two random walkers each starting from one of the nodes considered, along paths of possibly unequal length ($SimRank$) (Zhang et al. 2012).

As already mentioned in the previous section, when the objects under consideration are represented in a hierarchical manner, set intersectional similarity measures cannot capture this aspect. It can result in 0 similarity value between the objects of different heights in the hierarchy of concepts even though they might actually be similar.

Other types of similarity measures are vector space methods which include cosine similarity and Pearson Correlation Coefficient. A user study was conducted to evaluate these similarity measures and it was found that the similarity measure introduced by the authors gives results that are very close to human judgment. A performance-based comparison of six structural collaborative measures of similarity with Cosine Index and Pearson Correlation Coefficient is detailed in (Zhang et al. 2010). The results on two datasets: MovieLens and Netflix indicates that Salton Index, Jaccard Index, and Sorensen Index always have good performance. Cosine similarity produces good results as well. However, its computational complexity is very high to be applied to very large data.

A simple group-based similarity measure, $GroupRem$, defined on movie tags and popularity was defined in (Pera and Ng 2013). When compared with three most popular collaborative filtering techniques, $GroupRem$ outperformed them with respect to the Discounted Cumulative Gain (Croft, Metzler, and Strohman 2010).

A comparative study of similarity measures between binary vectors, which the authors call *binary similarity measures*, is described in (Choi, Cha, and Tappert 2010), where both negative and positive matches have been studied. Seventy six binary similarity measures are clustered (using hierarchical clustering) and evaluated according to the relationships between them.

Inspired by $PageRank$, $PageSim$ (Lin, King, and Lyu

Table 5: Comparison between similarity measures

Similarity Measure	Time	Space
SimRank	$O(Kn^2d^2)$	$O(n^2)$
Improved SimRank	$O(k^4n^2)$ $(k \leq n)$	$k^2 \times n^2$
PageSim	$O(C^2)$, $C = kr$	*O(Cn), C = kr*
E-Rank	$O(n^3)$, but more extensive evaluation to be considered in future work	future work
SimFusion	$O(Kn^2d)$, where d is the number of iterations	$O(n^2)$, where n is the total number of objects
P-Rank	$O(Kn^2d^2)$	$O(n^2)$
FriendTNS	*0.012*sec, for*N=1000, k=10*	

Table 6: Comparison between similarity measures

Similarity Measure	Compared with	Dataset used
SimRank	Co-citation (Small 1973)	Research Index[1]
Improved SimRank	SimRank	DBLP[2]; Image data (querying Google Image Search); Wikipedia[3]
PageSim	SimRank; Cosine TFIDF as a ground truth	crawled Webpages[4]
E-Rank	Enriches P-Rank by considering both in- and out-links	Enron Email dataset[5], Citation Network[6], DBLP[7]
SimFusion	SimRank (detailed description) and tf $\times$ idf	Search click through log
P-Rank	Extends SimRank	Synthetic[2]
Vertex similarity [2]	Cosine similarity and SimRank	AddHealth data: study as part of the National Longitudinal Study of Adolescent Health
FriendTNS	RWR, Shortest Path, Adamic/Adar, FOAF	Facebook, Hi5, Epinion

[1]http://www.researchindex.com Transcript of 1050 students at Stanford University;
[2]http://kdl.cs.umass.edu/data/dblp/dblp-info.html;
[3]http://www.wikipedia.org/
[4]http://www.cse.cuhk.edu.hk
[5]http://www.cs.cmu.edu/enron/
[6]http://snap.stanford.edu/data/
[7] http://www.informatik.uni-trier.de/ley/db/

2006) is a method for finding similar web pages in domains such as search engines or web document classifications, and it was evaluated against Cosine TF/IDF.

The Facebook "People you may know"friends recommender uses friends of friends, paths of length two, as a similarity measure and global graph properties (as local graph information) are used to recommend friends. Precision and recall were used to measure the performance of friend recommendation (Symeonidis, Tiakas, and Manolopoulos 2010), (Papadimitriou, Symeonidis, and Manolopoulos 2012).

Conclusion

Similarity measures play an important role in information porcessing. When used in conjunction with social networks (or more generally, complex networks) two main issues arise, structural, that is, the link pattern of the network, and semantic, that is, the meaning of nodes (the information stored in them). These issues led researchers to develop structural, semantic, and hybrid structural and semantic measures of similarity for such networks. This brief survey illustrates the variety of similarity measures developed for social networks and highlights the difficulty of selecting a similarity measure for problems such as link prediction or community detection. Tables 5 and 6, list the differences between several similarity measures. Table 5 compares the selected similarity measures based on time and space complexities, while Table 6 compares the selected similarity measures with respect to whom they compared their work with, dataset used, and performance. For a comparison between similarities from other perspectives the reader is referred to (Lin 1998) and (Choi, Cha, and Tappert 2010) and references therein.

References

Adamic, L. A., and Adar, E. 2003. Friends and neighbors on the web. *Social networks* 25(3):211–230.

Bahl, L. R.; Jelinek, F.; and Mercer, R. 1983. A maximum likelihood approach to continuous speech recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* (2):179–190.

Bhattacharyya, P.; Garg, A.; and Wu, S. F. 2011. Analysis of user keyword similarity in online social networks. *Social network analysis and mining* 1(3):143–158.

Burkhard, H.-D., and Richter, M. M. 2001. On the notion of similarity in case based reasoning and fuzzy theory. In *Soft computing in case based reasoning*. Springer. 29–45.

Choi, S.-S.; Cha, S.-H.; and Tappert, C. C. 2010. A survey of binary similarity and distance measures. *Journal of Systemics, Cybernetics and Informatics* 8(1):43–48.

Croft, W. B.; Metzler, D.; and Strohman, T. 2010. *Search engines: Information retrieval in practice*. Addison-Wesley Reading.

Díaz, I., and Ralescu, A. 2012. Privacy issues in social networks: a brief survey. In *Advances in Computational Intelligence*. Springer. 509–518.

Elavarasi, S Anitha, A. J., and Menaga, K. 2014. A survey on semantic similarity measure. *International Journal of Research in Advent Technology* 2(4):389–398.

Facebook. 2010. Facebook: 10 years of social networking, in numbers. [Online; accessed 19-July-2014].

Findler, N. V., and Van Leeuwen, J. 1979. A family of similarity measures between two strings. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* (1):116–118.

Ganesan, P.; Garcia-Molina, H.; and Widom, J. 2003. Exploiting hierarchical domain structure to compute similarity. *ACM Transactions on Information Systems (TOIS)* 21(1):64–93.

Heidemann, J.; Klier, M.; and Probst, F. 2012. Online social networks: A survey of a global phenomenon. *Computer Networks* 56(18):3866–3878.

Huang, X., and Lai, W. 2006. Clustering graphs for visualization via node similarities. *Journal of Visual Languages & Computing* 17(3):225–253.

Ilakiya, P.; Sumathi, M.; and Karthik, S. 2012. A survey on semantic similarity between words in semantic web. In *Radar, Communication and Computing (ICRCC), 2012 International Conference on*, 213–216. IEEE.

Jaccard, P. 1912. The distribution of the flora in the alpine zone. 1. *New phytologist* 11(2):37–50.

Jeh, G., and Widom, J. 2002. Simrank: a measure of structural-context similarity. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 538–543. ACM.

Jiang, J. J., and Conrath, D. W. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv preprint cmp-lg/9709008*.

Leicht, E.; Holme, P.; and Newman, M. E. 2006. Vertex similarity in networks. *Physical Review E* 73(2):026120.

Li, Y.; Bandar, Z. A.; and McLean, D. 2003. An approach for measuring semantic similarity between words using multiple information sources. *Knowledge and Data Engineering, IEEE Transactions on* 15(4):871–882.

Li, C.; Han, J.; He, G.; Jin, X.; Sun, Y.; Yu, Y.; and Wu, T. 2010. Fast computation of simrank for static and dynamic information networks. In *Proceedings of the 13th International Conference on Extending Database Technology*, 465–476. ACM.

Li, C. H.; Yang, J. C.; and Park, S. C. 2012. Text categorization algorithms using semantic approaches, corpus-based thesaurus and wordnet. *Expert Systems with Applications* 39(1):765–772.

Lin, Z.; King, I.; and Lyu, M. R. 2006. Pagesim: A novel link-based similarity measure for the world wide web. In *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, 687–693. IEEE Computer Society.

Lin, D. 1998. An information-theoretic definition of similarity. In *ICML*, volume 98, 296–304.

Lü, L., and Zhou, T. 2011. Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications* 390(6):1150–1170.

Mabotuwana, T.; Lee, M. C.; and Cohen-Solal, E. V. 2013. An ontology-based similarity measure for biomedical data–application to radiology reports. *Journal of biomedical informatics* 46(5):857–868.

Naderi, H., and Rumpler, B. 2007. Three user profile similarity calculation (upsc) methods and their evaluation. In *Signal-Image Technologies and Internet-Based System, 2007. SITIS'07. Third International IEEE Conference on*, 239–245. IEEE.

Pan, Y.; Li, D.-H.; Liu, J.-G.; and Liang, J.-Z. 2010. Detecting community structure in complex networks via node similarity. *Physica A: Statistical Mechanics and its Applications* 389(14):2849–2857.

Papadimitriou, A.; Symeonidis, P.; and Manolopoulos, Y. 2012. Fast and accurate link prediction in social networking systems. *Journal of Systems and Software* 85(9):2119–2132.

Pedersen, T.; Patwardhan, S.; and Michelizzi, J. 2004. Wordnet:: Similarity: measuring the relatedness of concepts. In *Demonstration Papers at HLT-NAACL 2004*, 38–41. Association for Computational Linguistics.

Pera, M. S., and Ng, Y.-K. 2013. A group recommender for movies based on content similarity and popularity. *Information Processing & Management* 49(3):673–687.

Pirró, G. 2009. A semantic similarity metric combining features and intrinsic information content. *Data & Knowledge Engineering* 68(11):1289–1308.

Rawashdeh, A.; Rawashdeh, M.; Díaz, I.; and Ralescu, A. 2014. Measures of semantic similarity of nodes in a social network. In *Information Processing and Management of Uncertainty in Knowledge-Based Systems*, 76–85. Springer.

Small, H. 1973. Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for information Science* 24(4):265–269.

Symeonidis, P.; Tiakas, E.; and Manolopoulos, Y. 2010. Transitive node similarity for link prediction in social networks with positive and negative links. In *Proceedings of the fourth ACM conference on Recommender systems*, 183–190. ACM.

Xi, W.; Fox, E. A.; Fan, W.; Zhang, B.; Chen, Z.; Yan, J.; and Zhuang, D. 2005. Simfusion: measuring similarity using unified relationship matrix. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, 130–137. ACM.

Xiao, J.-T. 2012. An efficient web document clustering algorithm for building dynamic similarity profile in similarity-aware web caching. In *Machine Learning and Cybernetics (ICMLC), 2012 International Conference on*, volume 4, 1268–1273. IEEE.

Yang, X.; Tian, Z.; Cui, H.; and Zhang, Z. 2012. Link prediction on evolving network using tensor-based node similarity. In *Cloud Computing and Intelligent Systems (CCIS), 2012 IEEE 2nd International Conference on*, volume 1, 154–158. IEEE.

Yang, X.; Guo, Y.; Liu, Y.; and Steck, H. 2014. A survey of collaborative filtering based social recommender systems. *Computer Communications* 41:1–10.

Zhang, Q.-M.; Shang, M.-S.; Zeng, W.; Chen, Y.; and Lü, L. 2010. Empirical comparison of local structural similarity indices for collaborative-filtering-based recommender systems. *Physics Procedia* 3(5):1887–1896.

Zhang, M.; He, Z.; Hu, H.; and Wang, W. 2012. E-rank: A structural-based similarity measure in social networks. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2012 IEEE/WIC/ACM International Conferences on*, volume 1, 415–422. IEEE.

Zhao, P.; Han, J.; and Sun, Y. 2009. P-rank: a comprehensive structural similarity measure over information networks. In *Proceedings of the 18th ACM conference on Information and knowledge management*, 553–562. ACM.

Zhou, Y.; Cheng, H.; and Yu, J. X. 2009. Graph clustering based on structural/attribute similarities. *Proceedings of the VLDB Endowment* 2(1):718–729.