Third International Conference
Modelling and Development of Intelligent Systems
October 10 - 12, 2013
"Lucian Blaga" University
Sibiu - Romania

Towards a Unified Similarity Measure for Node Profiles in a Social Network

Ahmad Rawashdeh and Anca Ralescu

Abstract

Assessing the similarity between node profiles in a social network is an important tool in its analysis. Several approaches exist to study profile similarity, including semantic approaches and natural language processing. However, to date there is no research combining these aspects into a unified measure of profile similarity. Traditionally, semantic similarity is assessed using keywords, that is, formatted text information, with no natural language processing component. This study proposes an alternative approach, whereby the similarity assessment based on keywords is applied to the output of natural language processing of profiles. A *unified similarity measure* results from this approach. The approach is illustrated on a real data set extracted from Facebook.

1 Introduction

Social networks allow people to connect and share their personal details. Many social networking websites have been created and they vary in the services which they provide. Mainly, they allow users to comment and post pictures or video and share.

Facebook is a social networking website that has over one billion users. It allows the user to connect to friends, create personal profiles by specifying their interest –TV, movies, sports, and books – and by posting images and videos of their activities. The website also allows anyone to create pages for their business or favorite personality. Users can even create pages for special interest groups which are open on a restricted basis to group members.[6]

People tend to form relationships with people who are similar to them. Alternatively, it can be said that if a relationship is formed between two people, then there must be some similarity between them. Indeed, it has been found that 80% of social network users form relationships with the contact of their friends [3].

Analysis of similarity between Facebook profiles can be assessed from the study of keyword similarity [3]. To find the relationship between the keywords, these are arranged in a hierarchical structure to form trees of different heights. In the forest model more than one tree is generated for each profile. Related words are retrieved by search in these profile trees, implemented as heuristic search. Semantic relationships between the words can be assessed by using Wordnet. [10]

This study proposes to find the semantic relationship between attribute entries in the social network, not only between keywords. Therefore the category of the words which appear in these entries must be found. This can be accomplished by using a tagger, a program which tags a word by its semantic category. These categories are used to extract the words suitable to assess profile similarity [4]. The (semantic) distance between profiles is very important to this process, as it has been shown that the similarity between profiles deteriorates as the distance between them increases [4].

From this point on, the paper is organized as follows: Section 2 describes the proposed approach for similarity assessment. Section 3 presents the data and the results obtained from applying this approach on a Facebook data-set. The paper closes with a discussion and conclusion section.

2 Finding Similar Profiles

The measure of similarity proposed here combines Wordnet [8] and cosine similarity, which is a very common device to assess document similarity [9].

2.1 Wordnet

Wordnet is a free lexical database that organizes English words into concepts and relations, well-known for assessing semantic similarity. English nouns, verbs, adjectives, and adverbs form hierarchies of *synset* where relations exist that connect them. The relations are Synonymy, Antonymy, Hypernymy, Meronymy, Troponymy, Entailment.

Hypernym of a word

Hypernym of a word conveys its place in a hierarchy of concepts/words and can be retrieved using Wordnet. Consider for example, the two senses of word

”comedy”:

- comedy as a ”humorous drama”
- comedy as ”comic incident”

Taking the first sense, since comedy is kind of drama, drama is a hypernym of comedy. Similarly, since drama is kind of literary work, literary work is a hypernym of drama [5]. The hierarchy determined by the hypernym relationship is a *synset*. Therefore, based on the above, the synset for comedy (with respect to the first meaning) is

Synset 1: [entity] ← [abstract entity] ← [abstraction] ← [communication] ← [expressive style,style] ← [writing style,literary genre,genre] ← [drama] ← [comedy] - light and humorous drama with a happy ending (1)

while the Synset with respect to the second meaning is:

Synset 2: [entity] ← [abstract entity] ← [abstraction] ← [communication] ← [message,content,subject matter,substance] ← [wit, humor, humor, witticism, wittiness] ← [fun, play,sport] ← [drollery, clowning, comedy, funniness] - a comic incident or series of incidents (2)

2.2 Cosine Similarity

Cosine similarity [9] has been successfully used as measure of similarity between documents. A document is described by a vector of fixed dimension of word frequencies. The similarity of two documents is assessed based on the cosine of the angle made between their corresponding vectors. More precisely, given the documents D_i, $i = 1, 2$, with corresponding word vectors v_1 and v_2, the *cosine similarity* between D_1 and D_2 and d2 is defined as

$$CS(D_1, D_2) = \frac{v_1 \cdot v_2}{\|v_1\|\|v_2\|} \tag{3}$$

where $\cdot$ is the dot product between two vectors, and $\|v\|$ denotes the norm of a vector v. Table 1 shows the result of evaluating the cosine similarities between three documents with associated vectors given as

$$\begin{array}{rcl} v_1 & = & (2, 3, 5, 10) \\ v_2 & = & (6, 2, 3, \ 0) \\ v_3 & = & (0, 1, 2, \ 0) \end{array} \tag{4}$$

Table 1: Cosine similarity between the vectors v_1, v_2, and v_2 of (4).

	v_1	v_2	v_3
v_1	1.0000	0.4013	0.4949
v_2	0.4013	1.0000	0.5111
v_3	0.4949	0.5111	1.0000

As it can be seen from this table, the largest cosine similarity is between the 2nd and 3rd document, followed by that between 1st and 3rd document. This corresponds to the first two smallest distances between the vectors v_2 and v_3, and v_1 and v_3 (and will always be so, since the vectors have positive components).

2.3 A Unified Similarity

The approach for the is illustrated on the computation of the similarity between two Facebook profiles. The following steps are performed:

1. Extract the text in the feature field (movies, title) if the data-set is not formatted well.
2. Natural Language Processing: Parse the sentence to obtain its structure.
3. Get the first synset of the word using Wordnet.
4. Encode the word
   - Get all hypernym of the synset of the word.
   - Find the distance from the word to the root of the synset.
5. Each feature field of a profile is encoded as a vector of such distances.
6. Apply cosine similarity between vectors of such distances.

The NLP component in step 2, is used to label (tag) words according to their speech category [7]. The categories used in this study are: NN (noun, proper, singular or mass), NNP (noun, proper, singular), NNS (noun, common, plural), and NNPS (noun, proper, plural) [1]. These part of speech tags are used to assess profile similarity.

The innovative aspect of the current approach is in the encoding of the text input into a vector of distances. This is done as follows: For each profile, the outcome of Step 2 is a collection of word-tag pairs (w, t_w). Given a word-tag pair, (w, t_w), w is considered for inclusion in the similarity evaluation if and only if $t_w \in Tags$, where $Tags = \{NN, NNS, NNP, NNPS\}$ denotes a set of tags of interest. Next, each selected word, w is input to Wordnet which returns the list of hypernyms, in the hierarchical synset representation of w. As illustrated in the example above on the word "comedy" more than one synset can be returned by Wordnet. In this study, only the first sysnset is used for similarity assessment. The encoding of w is the distance to it from the top hypernym ('entity') in the synset. For example, the encoding of the word "comedy" based on the first synset 1 is equal to 7.

If a word has no hypernym (e.g., it is not in Wordnet) then its encoding is 0. This process is summarized as follows. Represent a profile p as a vector of words. That is, $p = [w_1, \ldots, w_k]$ where w_i, $i = 1, \ldots, k$ is a word extracted from the profile by the tagger and k is the number of words extracted.

For each word w_i, use Wordnet to extract its first synset. Define $d_i = d(w_i)$ where, for a given word w,

$$d(w) = \begin{cases} dist(w, \text{[entity]}) & \text{if } w \text{ is in Wordnet} \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where $dist$ is the distance to [entity], the top hypernym of w in its first synset, output by Wordnet. The encoding of the profile p is a mapping $e : p \mapsto \Re_+^k$ such that

$$e(p) = (d_1, \ldots, d_k)$$

Given two profiles, p, and p' and their corresponding encoding $e(p) = (d_1, \ldots, d_k)$ and $e(p') = (d'_1, \ldots, d'_k)$ the similarity between p and p' is defined as the cosine similarity of $e(p)$ and $e(p')$, as shown in equation (6)

$$Sim(p, p') = CS(e(p), e(p')) \tag{6}$$

where CS is defined as in equation (3).

The process described above converts the problem of similarity assessment between unstructured data into a more rigorously defined problem of similarity between real valued vectors. In principle, it is possible, for a given word w (and hence for a profile), to obtain more than one encoding, by using all the synsets to encode a line of text using several synsets. However, this case is beyond the scope of the current study. Figure 1 illustrates the approach proposed in this study and described above.

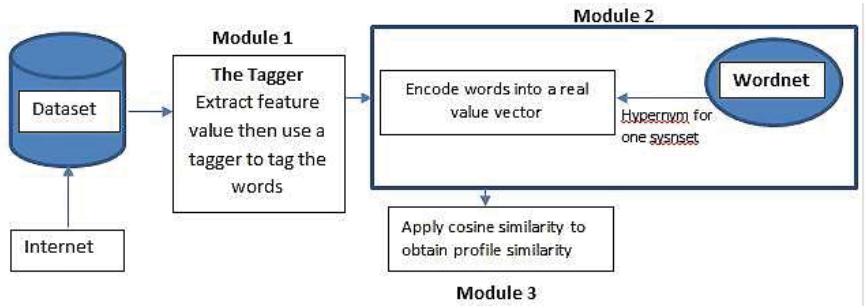

Figure 1: Diagram for computing the unified similarity measure.

The Occurrence Frequency Similarity (OF) of Node Profiles

Let u and x denote two profiles, each having the multiple valued attribute i. The *occurrence frequency* similarity measure, OF, between u and x is defined by equation (7) following the work in [2]. This measure of similarity will be used for comparisons with the measure proposed in this study.

$$OF(i_u, i_x) = \frac{1}{B}\sum_{k=1}^{B}\begin{cases} 1 & \text{if} \quad i_u.n = i_x.n \\ (1 + A \times B)^{-1} & \text{if} \quad i_u.n \neq i_x.n \end{cases} \tag{7}$$

where B is the number of attributes, i_u and i_x are the values of attribute i in the profiles u and x respectively, $i_u.n$ and $i_x.n$ denote the value of the nth subfield for i_u and i_x respectively, N is the total number of item values, and $f(\cdot)$ is the number of records; $A = \log(\frac{N}{1+f(i_u.n)})$, and $B = \log(\frac{N}{f(i_x.k)})$.

3 Experimental Results

The approach described in the preceding section is applied to a Facebook data set as shown next.

3.1 Facebook Profiles Data-set

The Facebook data-set considered in experiments contains 2013 profile pages from Facebook (raw data before the introduction of the Facebook time-line). Skull security has a list of publicly available Facebook URLs which is used to download this data-set that consists of 2013 profiles [2]. More specifically, *Data-set.txt* (Facebook Data-set) contains all the movies interest for different Facebook profile numbers. The format of the data-set is as follows: *Profile_id* followed by the Movies interest entered by the user identified by the *Profile_id*. Furthermore, various characteristics are extracted from the Facebook Data-set, as shown in Table 2. Figure 2 shows the frequency of the top 20 movies in the

Table 2: Characteristics of the Facebook profile data.

Number of Facebook profiles	2013
Average movies entries per profile	2.9
Number of movies entries for all profiles	1744
Maximum movies entries	8
Most Common Genre type [1]	which is the genre type ”unknown”
Minimum movies entries	0
Different movies count	1089

Facebook data-set.

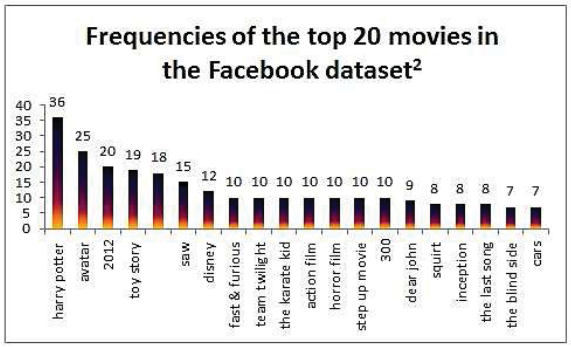

Figure 2: Frequency of the top 20 movies from the Facebook data-set.

Table 3 illustrates the encoding the Movie Attribute for three Facebook pro-

files.

Table 3: Illustration of Movie Attribute of Facebook profiles: their tags and Hypernyms.

Profile 1: Movie Attribute	Harry Potter, Transformers, Mr. & Mrs. Smith						
Words	Harry	Potter	Transformers	Mr.	&	Mrs.	Smith
Tags	NNP	NNP	NNPS	NNP	CC	NNP	NNP
dist to root in synset	0	7	8	8	ignored	8	0
Profile 2: Movie Attribute	Sherina's Adventure						
Words	Sherina	's	Adventure				
Tags	NNP	POS	NNP				
dist to root in synset	0	ignored	8				
Profile 3: Movie Attribute	Love mein Gum, Maqsood Jutt Dog Fighter						
Words	Love	mein	Gum	Maqsood	Jutt	Fog	Fighter
Tags	NNP	NNP	NNP	NNP	NNP	NNP	NNP
dist to root in synset	7	0	7	0	0	6	4

3.2 Results

The algorithm of [4] and the approach described here were implemented in Java. The similarity was calculated between each adjacent nodes' line in the data-set using both the OF measure and Wordnet approach. As we can see from the results, since the Occurence Frequency (OF) depends on whether or not there are redundant data in the data-set. Table 4 illustrates these similarity results for two profiles using OF and Wordnet approaches.

Table 4: OF and Wordnet Similarity of two Facebook profiles along their Movie Attribute.

Data Set	Facebook
Profile-1 ID	100000060663828.html
Movies Interests	Captain Jack Sparrow, Meet The Spartans, Ice Age Movie, Spider-Man
Profile-2 ID	100000067167795.html
Movies Interests	Clash of the Titans, Ratatouille, Independence Day, Mr. Nice Guy, The Lord of the Rings Trilogy (Official Page)
OF Similarity	0.9472
Wordnet based similarity	0.1892

Figure 3 shows the result of applying the OF algorithm to find the similarity and the semantic Wordnet based method for all the node pairs connected by an edge in the data set. Using OF, most of the data are similar, with similarity

value equal to 1. But using Wordnet, the similarity values are distributed over all the data having a peak value at 0.2.

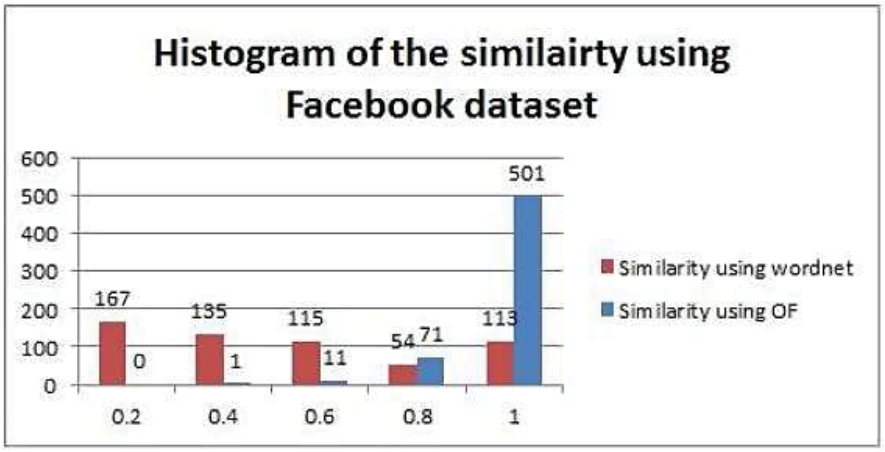

Figure 3: OF and Wordnet similarity results for the Facebook data-set.

4 Conclusions

This study introduces a new approach towards a unified measure of similarity between node profiles, and in general, between pieces of unstructured text. Natural language processing is used to extract speech parts from the texts of interest, and to encode them into vectors with positive components using the distance between the words extracted to the root of a hierarchy of concepts. Similarity is then evaluated between the resultant encoding vectors. While the results seem promising, several issues remain to be discussed and developed in subsequent studies.

References

[1] The university of pennsylvania (penn) treebank tag-set. `http://www.comp.leeds.ac.uk/ccalas/tagsets/upenn.html`. [Online; accessed 1-October-2013].

[2] Cuneyt Gurcan Akcora, Barbara Carminati, and Elena Ferrari. Network and profile based measures for user similarities on social networks. In *Information Reuse and Integration (IRI), 2011 IEEE International Conference on*, pages 292–298. IEEE, 2011.

[3] Prantik Bhattacharyya, Ankush Garg, and Shyhtsun Felix Wu. Analysis of user keyword similarity in online social networks. *Social network analysis and mining*, 1(3):143–158, 2011.

[4] Shyam Boriah, Varun Chandola, and Vipin Kumar. Similarity measures for categorical data: A comparative evaluation. *red*, 30(2):3, 2008.

[5] Ronald Bowes. Return of the Facebook Snatchers. `http://www.skullsecurity.org/blog/2010/return-of-the-facebook-snatchers`, 2010. [Online; accessed 19-July-2012].

[6] Nicole B Ellison et al. Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1):210–230, 2007.

[7] The Stanford Natural Language Processing Group. Pos Tagger FAQ. `http://nlp.stanford.edu/software/pos-tagger-faq.shtml`. [Online; accessed 19-July-2012].

[8] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[9] Helen J Peat and Peter Willett. The limitations of term co-occurrence data for query expansion in document retrieval systems. *JASIS*, 42(5):378–383, 1991.

[10] Matt Spear, Xiaoming Lu, Norman S Matloff, and S Felix Wu. Inter-profile similarity (ips): a method for semantic analysis of online social networks. In *Complex Sciences*, pages 320–333. Springer, 2009.

Ahmad Rawashdeh
Machine Learning and Computational Intelligence Lab
Department of Electrical Engineering and Computing Systems
University of Cincinnati, ML 0008
Cincinnati, OH 45221, USA
E-mail: *rawashmy@email.uc.edu*

Anca Ralescu
Machine Learning and Computational Intelligence Lab
Department of Electrical Engineering and Computing Systems
University of Cincinnati, ML 0008
Cincinnati, OH 45221, USA
Anca.Ralescu@uc.edu