Exploring Wikipedia and DMoz as Knowledge Bases for Engineering a User Interests Hierarchy for Social Network Applications

Mandar Haridas and Doina Caragea

Kansas State University
Nichols Hall, Manhattan, KS 66502
{mandar,dcaragea}@ksu.edu

Abstract. The outgrowth of social networks in the recent years has resulted in opportunities for interesting data mining problems, such as interest or friendship recommendations. A global ontology over the interests specified by the users of a social network is essential for accurate recommendations. We propose, evaluate and compare three approaches to engineering a hierarchical ontology over user interests. The proposed approaches make use of two popular knowledge bases, Wikipedia and Directory Mozilla, to extract interest definitions and/or relationships between interests. More precisely, the first approach uses Wikipedia to find interest definitions, the latent semantic analysis technique to measure the similarity between interests based on their definitions, and an agglomerative clustering algorithm to group similar interests into higher level concepts. The second approach uses the Wikipedia Category Graph to extract relationships between interests, while the third approach uses Directory Mozilla to extract relationships between interests. Our results show that the third approach, although the simplest, is the most effective for building a hierarchy over user interests.

1 Introduction

Over the years, there has been a dramatic increase in the number of social network users. Collectively, the top ten social network sites have grown at a rate of almost fifty percent every year [1]. In the coming years, it is expected that social networking will become more ingrained in mainstream sites. As a consequence, there is a great need for data mining in social networks. For example, using data mining, users can be recommended "new interests" based on the interests of their current friends. Similarly, users can be recommended "new friends" based on their current interests and friends. To address such social network problems effectively, it is essential to organize user interests into an ontology, in particular a hierarchical ontology. An ontology is an explicit formal specification of the terms and relations among terms in a domain [2]. It can be achieved by a systematic grouping of domain concepts (e.g., user interests) based on their definitions, in machine-interpretable form.

R. Meersman, T. Dillon, P. Herrero (Eds.): OTM 2009, Part II, LNCS 5871, pp. 1238–1245, 2009.

Constructing an ontology over the interests specified by the users of a social network has attracted attention among some researchers previously. Bahirwani et al. [3] have constructed an interest ontology by fetching the definitions of interests from three online sources, namely WordNet-Online, Internet Movie Database (IMDB) and Amazon Associates Web Services (AWS). Each definition of an interest can be seen as an instance. Similarity between instances is computed as the dot product of the vectors representing the instances. Instances are grouped into an ontology using a hierarchical agglomerative clustering approach [4].

Although the ontology constructed in [3] has proven helpful for improving the predictions of friendship relationships, the use of WordNet-Online, IMDB and AWS for a semantic understanding of user interests is cumbersome and may not always give complete and accurate definitions of interests. For interests belonging to topics, such as *Sports Persona*, that do not fit in the categories of *movies*, *books* and *words*, one would need to get definitions from a different knowledge base. For example, for the user interest *Pete Sampras*, definitions can be found neither on WordNet-Online nor on IMDB and AWS. Thus, the use of such discrete, distributed knowledge bases is inconvenient and undesirable, and cannot represent interests from diverse topics.

Furthermore, once the definitions corresponding to interests are obtained, the clustering approach in [3] groups the definitions in a binary tree hierarchy, as opposed to an n-ary tree. The resulting clusters do not capture the concept information for related interests. If two interests, e.g., *Laptops* and *Notebooks* are clustered, then the title of the new cluster becomes *Laptops & Notebooks*, when instead it would be desirable to derive the title *Portable Computers* to indicate the semantics of its children in a more meaningful way.

Our work explores different ontology engineering approaches and more comprehensive knowledge bases (in particular, Wikipedia and Directory Mozilla) to address the limitations mentioned above. A data set consisting of 1,000 users of the LiveJournal online service is used in this study. There are approximately 22,000 interests that these users have collectively specified. These interests belong to a wide variety of domains, including *Movies*, *Books*, *Sports*, *Social* and *Current Issues*, among others. In our first approach, we obtain definitions of interests from Wikipedia and use the technique of latent semantic analysis (LSA) to measure the similarity between interests. While this approach produces a more sensible ontology than the one produced by the approach in [3], this ontology is still a binary tree and consists of internal clusters labelled based on child information. Our second and third approaches explore the reuse of knowledge from existing hierarchies such as the Wikipedia Category Graph (WCG) and Directory Mozilla (DMoz), respectively, to group interests. For the implementation of the three approaches, Wikipedia dump from October 8, 2008 and DMoz dump from November 5, 2008 were used.

After discussing related work in Section 2, we briefly describe the three approaches considered in our work in Sections 3.1, 3.2., 3.3, respectively. More details can be found in [5]. We conclude the paper with a summary, short discussion and ideas for future work in Section 4.

2 Related Work

Exploiting the comprehensibility and coverage of Wikipedia has been a focus of widespread research. Amongst various other works, [6] and [7] have used Wikipedia to assign category labels to documents. Syed et al. [8] have used Wikipedia to predict categories common to a set of documents, whereas [9], [10], and [11] have used Wikipedia to compute semantic relatedness between documents. Furthermore, the LSA technique has been employed for text categorization in previous work. As an example, Lee et al. [12] have used LSA for multi-language text categorization. Our Wikipedia/LSA approach makes use of the advantages that Wikipedia, as a data source, and LSA, as a text categorization technique, have to offer when engineering an ontology of user interests. The usefulness of DMoz for classification problems have been previously demonstrated as well. Grobelnik and Mladeni [13] have shown that the use of large topic ontologies such as DMoz can help in classifying Web documents.

As opposed to the previous work, the contribution of our work lies in the fact that we construct an accurate user interest n-ary hierarchy by efficiently and effectively reusing a single comprehensive knowledge base that covers a wide variety of interest topics. In the process, we derive useful observations about the effectiveness of the LSA technique to compute the similarity between interests. Also, we compare the usefulness of Wikipedia versus DMoz as data sources in the ontology engineering process.

3 Proposed Hierarchy Engineering Approaches

3.1 Wikipedia/LSA Approach

In this approach, we obtain definitions of interests from Wikipedia and compare the definitions using the standard LSA technique [14]. For illustration purposes, we will use a small interests set consisting of 10 interests, picked up at random, from diverse fields such as *Movies*, *Sports* and *Current affairs*. The hierarchy constructed using the approach in [3] is shown in Figure 1. In the figure, nodes 0 to 8 indicate the order in which the user interests are clustered, with 0 indicating the first clustering and 8 indicating the last clustering. As can be seen, this approach does not perform very well for the set of interests considered. For example, user interests *Tom Hanks* and *9-11* are clustered first. We will compare the ontology produced with the Wikipedia/LSA approach with this baseline.

To start with, for every user interest in our data, the relevant Wiki document is fetched (and regarded as the definition of the corresponding interest). The fetched documents are cleaned, as follows: text in the documents is tokenized, tokens are stemmed and stop words are removed. From the tokenized documents, a term-document matrix is constructed. The term-document matrix is then decomposed using singular value decomposition (SVD) technique [15] and the dimensionality of the decomposed matrix is reduced. The advantage of

such reduction is that interest documents that have many words in common get grouped closer to each other. Hidden relationships between interest documents are discovered, while weak undesired relationships get eliminated.

After applying the LSA technique, each document is represented as a vector of weights [16]. Similarity between a pair of documents is computed as the cosine of the angle between the corresponding document vectors. After computing the cosine similarity between each pair of interest documents in our data set, we then cluster the documents using a hierarchical agglomerative clustering algorithm [4].

Figure 2 shows the ontology that is constructed using this approach. User interests *age of empires* and *computer gaming* are clustered first (to form a new node 0), as they have the highest similarity. Next, user interests *Tom Hanks* and *Oscar Wilde* are clustered (to form a new node 1), as they have the next highest similarity measure, and so on. While not perfect, the resulting ontology is more accurate than the ontology engineered using the approach in [3]. The

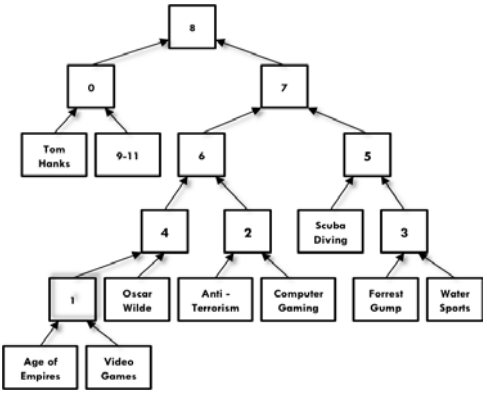

Fig. 1. Hierarchy over a set of 10 user interests when WordNet/IMDB/AWS are used

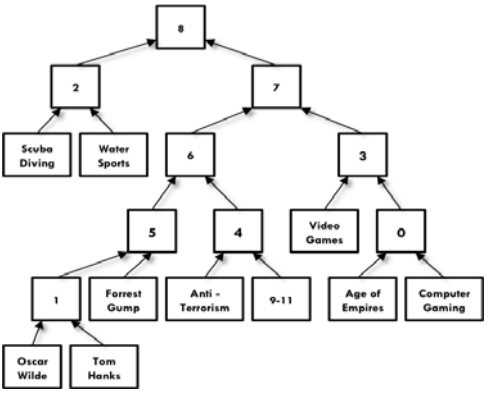

Fig. 2. Hierarchy over the set of 10 user interests using the Wikipedia/LSA approach

use of Wikipedia to obtain interest definitions results in good coverage for users interests, without the need for multiple sources such as WordNet-Online, IMDB, etc., as explained in Section 1. Furthermore, the use of LSA helps in reducing the noise in the data and unhiding latent relationships between documents.

However, the approach has several shortcomings. First, the ontology engineered is still a binary tree. Second, new nodes that are constructed as a result of clustering of interest instances do not have explicit semantics associated with them. Therefore, we explore two more approaches to engineer the ontology.

3.2 WCG Based Approach

In Wikipedia, every article belongs to some category. The articles form a network of semantically related terms, while categories are organized in a taxonomy-like structure, called WCG [17]. In the WCG based approach, we exploit the category information contained within the WCG. Just as in the Wikipedia/LSA approach, we obtain definitions of interests from Wikipedia. However, instead of comparing the documents corresponding to interests, we compare categories corresponding to interest documents. Now, interests belonging to the same categories are grouped together. This is done by mining the WCG for relationships between categories and grouping the categories themselves with each other.

As an example, the interest *Tom Hanks* belongs to categories *Film Actors*, *1956 Births*, *Best Actor Oscars*, among others. Nodes for the three categories are created with the *Tom Hanks* node as child for each category. Furthermore, the category *Best Actor Oscars* is a subcategory of the category *Film Actors*. Therefore, the former is appended as a child of the later. Such relationships between categories are extracted from the WCG.

The WCG based approach serves the purpose of assigning semantics to newly formed clusters in the hierarchy. However, in Wikipedia, each article can belong to an arbitrary number of categories. The large number of categories results in large scale duplication of interest instances in the resulting hierarchy. Furthermore, WCG contains cycles and disconnected categories [17]. Breaking the cycles requires further duplication of nodes. Another drawback of this approach is that it is not possible to directly extract the complete category link information for a Wikipedia article. Due to the above mentioned shortcomings, this approach fails to provide the desired results. However, it motivates our next approach: if we can extract the complete category link to which interests belong and if we can rank the categories to which the interests belong based on their importance, the interests can be classified effectively and grouped into a hierarchy.

3.3 DMoz Based Approach

The problem of retrieving the "complete category link" for an interest and ranking the categories based on their importance is resolved through the use of DMoz category hierarchy. Every category in the DMoz dump consists of a listing and description of external pages associated with that category.

Our DMoz based approach works as follows: each interest is searched in the DMoz RDF dump. The categories under which one or more of the external page descriptions contain the concerned interest are selected. Thus, when searching for an interest such as *9-11* in the dump, we find that *9-11* occurs frequently in the external page description of the category link *Society* → *Issues* → *Terrorism* → *Incidents*. Similarly, *9-11* is also found (at least once) under an external page description of the category link *Arts* → *Movies* → *Titles* → *Fahrenheit 9-11*. All such category links under which the interest is found are extracted. The selected categories are ranked in decreasing order of the matches. To engineer the ontology, we use only the top five ranked categories. Thus, unlike with the WCG, for a very commonly occurring term which may belong to multiple categories, we prevent a large number of categories being engineered in the ontology by considering only the top five categories. This avoids large scale duplication of the interest instances. Furthermore, with DMoz, it becomes possible to retrieve the complete category link associated with the interest. Thus, for example, the interest *9-11* in DMoz belongs not only to the category *Incidents* but to its complete category link, which is *Society* → *Issues* → *Terrorism* → *Incidents*.

We parse the complete category link and every term in the link becomes a node in the ontology. The interest is made a child of the lowest node in the hierarchy. Figure 3 shows a fragment of the hierarchy constructed for a set of interests (*Tom Hanks*, *Brad Pitt*, *Sean Penn*, *Ballet*, *Michael Jackson*, *Cindrella*, *Troy*, *Forrest Gump*) under the concept *Arts*. As seen in the figure, all interests are accurately grouped under the respective categories to which they belong. The hierarchy constructed from all 22,000 interests in our data has 14 levels, 68281 leaf nodes (a consequence of multiple meanings for the interest words) and 52106 internal nodes. The maximum number of children for a node is 1912 (for the node "bands and artists", a sub-category of music). We should note that data mining techniques could be used to perform interest-word sense disambiguation, for example, by exploiting other interests of a user and the interests of his or her friends. Thus, as can be seen from Figure 3, this approach constructs a simple yet effective grouping of user interests. It addresses all the issues discussed in Section 1, presenting several advantages over the other approaches considered.

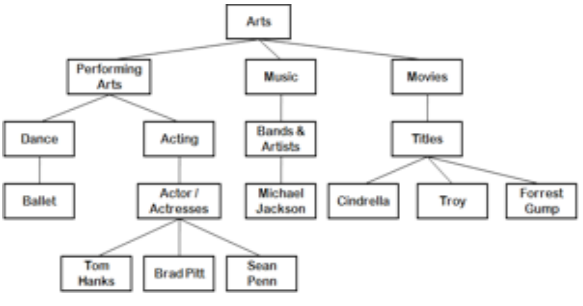

Fig. 3. Fragment of hierarchy engineered using DMoz

4 Summary, Discussion and Future Work

In this study, we have explored three approaches to the problem of building an ontology over the interests specified by the users of a social network. The first and third approaches produced usable hierarchies, although the Wikipedia/LSA hierarchy presents some limitations. While the second approach did not produce a useful ontology, it served as a bridge between the Wikipedia/LSA approach and DMoz approach. Moreover, it motivated the reuse of knowledge from existing hierarchies in the ontology engineering process.

Although the Wikipedia/LSA approach gives good results, it is computationally expensive (time and memory). Furthermore, our study shows that Wikipedia articles are detailed, but not always precise. With DMoz, the opposite is true. Its category hierarchy is "crisp." Searching for a term in the DMoz dump enables finding precise and accurate information as far as classifying the term is concerned, even if the approach used is very simple. This is not surprising, however, because the simplicity of the approach is compensated by the rich categorization that DMoz provides. Similar to Wikipedia, the DMoz category hierarchy covers a wide variety of topics ranging from *Arts*, *Sciences*, *Computers* to *Movies*, *Business*, *Health*, etc. This range of topics covers most of the domains, as far as user interests are concerned. Very importantly, the ontology engineered addresses all issues raised in Section 1.

In summary, while other authors such as Gabrilovich and Markovitch [10] have found that Wikipedia is better than DMoz for certain tasks (e.g., computing semantic relatedness), our study shows that in the case of social network user interests, DMoz serves better than Wikipedia in the ontology engineering process. Thus, one cannot claim Wikipedia to be better than DMoz, in general. The vice-versa is also true. Although our study shows that the DMoz hierarchy can help engineer better ontologies for most domains (as far as "interests" are concerned), this may not always be true. For example, for certain domains such as *Bioinformatics*, the DMoz hierarchy may not have adequate coverage. In such cases, a Wikipedia-based approach may be needed.

As shown above, the Wikipedia/LSA approach produces a potentially useful ontology over the interest documents. We plan to improve this approach by combining it with an approach to predict the concept associated with a group of documents [8], and thus associate semantics with the clusters formed. Furthermore, a slicing algorithm [18] will be used to transform binary hierarchies into n-ary hierarchies. An extensive exploration of the usefulness of both Wikipedia/LSA and DMoz-based interest hierarchies for the task of predicting friendship links is also part of our future work plans.

Acknowledgements

This work is supported by the National Science Foundation under Grant No. 0711396 to Doina Caragea. We would like to thank Dr. William H. Hsu and the KDD group at K-State for sharing their LifeJournal data with us.

References

1. Bausch, S., Han, L.: Social networking sites grow 47 percent, year over year, reaching 45 percent of web users, according to nielsen/netratings (2006), http://www.nielsen-online.com/pr/pr$_$060511.pdf
2. Gruber, T.: A translation approach to portable ontology specifications. Technical report 5(2), 199–220, Knowledge Systems AI Laboratory, Stanford University (1993)
3. Bahirwani, V., Caragea, D., Aljandal, W., Hsu, W.: Ontology engineering and feature construction for predicting friendship links in the LiveJournal social network. In: The 2nd SNA-KDD Workshop 2008, Las Vegas, Nevada, USA (2008)
4. Jardine, N., van Rijsbergen, C.J.: The use of hierarchical clustering in information retrieval. Information Storage and Retrieval 7, 217–240 (1971)
5. Haridas, M.: Exploring Wikipedia and DMoz as knowledge bases for engineering a user interest hierarchy for social network applications. M.S. Thesis, Department of Computing and Information Sciences. KSU, Manhattan, KS, USA (2009)
6. Gabrilovich, E., Markovitch, S.: Overcoming the brittleness bottleneck using Wikipedia: Enhancing text categorization with encyclopedic knowledge. In: The 21st National Conference on Artificial Intelligence, Boston, MA (2006)
7. Janik, M., Kochut, K.: Wikipedia in action: Ontological knowledge in text categorization. Technical report no. uga-cs-tr-07-001, University of Georgia (2007)
8. Syed, Z.S., Finin, T., Joshi, A.: Wikipedia as an ontology for describing documents. In: The 2nd International Conference on Weblogs and Social Media (2008)
9. Strube, M., Ponzetto, S.P.: WikiRelate! computing semantic relatedness using Wikipedia. In: The 21st National Conf. on AI, Boston, MA (2006)
10. Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In: The 20th International Joint Conference on Artificial Intelligence, Hyderabad, India (2007)
11. Milne, D.: Computing semantic relatedness using Wikipedia link structure. In: The New Zealand Computer Science Research Student Conference (2007)
12. Lee, C.H., Yang, H.C., Ma, S.M.: A novel multi-language text categorization system using latent semantic indexing. In: The First International Conference on Innovative Computing, Information and Control, Beijing, China (2006)
13. Grobelnik, M., Mladeni, D.: Simple classification into large topic ontology of web documents. In: The 27th International Conference on Information Technology Interfaces, Cavtat, Croatia (2005)
14. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. Journal of the American Society for Information Science 41, 391–407 (1990)
15. Manning, C., Raghavan, P., Schutze, H.: Introduction to Information Retrieval. Cambridge University Press, Cambridge (2008)
16. Rosario, B.: Latent semantic indexing: An overview. Final paper INFOSYS 240. University of Berkeley (2000)
17. Zesch, T., Gurevynch, I.: Analysis of the Wikipedia category graph for NLP applications. In: The TextGraphs-2 Workshop (2007)
18. Maarek, Y.S., Shaul, I.Z.B.: Automatically organizing bookmarks per contents. Comput. Netw. ISDN Syst. 28(7-11), 1321–1333 (1996)