Inferring dynamic taxonomies for terms based on UGC

Reem Alfayez and Mike Joy
Department of Computer Science
University of Warwick
Coventry, UK
r.qadan-al-fayez@warwick.ac.uk

Abstract– **Users of the web are currently the main content authors. Social networks form a valuable source of large volumes of user-generated content that might be beneficial in different areas of research. In this paper, we exploit the data generated by users of micro-blogging services, in particular Twitter, for disambiguating terms and inferring possible taxonomies for terms. We conducted an exploratory study to test the possibility of inferring high-level categories and possible sub-categories for which terms might be included. This experiment exploits the collection of hashtags which are mentioned with a specific hashtag in Tweets text., and the Open Directory Project (ODP), in order to discover dynamic taxonomies for a term (hashtag) with no knowledge needed about the semantic meaning of that term. We present several experiments which we have used to test this method, and promising results are reported.**

Index Terms– **Classification, Twitter, ODP, UGC, Relation discovery, Term disambiguation.**

I. Introduction

The recent emergence of social networks has increased popular engagement with the web, and web users are playing a vital role in authoring web content. Web 2.0 technologies, such as wikis, social networks and blogs, allow any ordinary user of the web to publish and share their ideas, views, and knowledge. User-generated content has been exploited in different areas of research such as business [5, 8], geographical information systems [4] and education [10].

Connecting content from the web on the fly is a massive capability for any system to have. In commerce, linking related products can help companies enhance online shopping by illustrating competitors' prices. In library management systems, linking related articles improves the usability of the system, and many other examples can be given. Relations between documents are established via links between key terms, labeled by authors, defined by a rigid classification, semantic analysis of meaning, or existing ontologies derived from expert knowledge, all of which require human effort in authoring document files and labeling them with the correct terms.

User-generated content extracted from social networks, and in particular, micro-blogging services such as Twitter, can be exploited to discover related terms and derive high level categories for collections of terms which frequently appear together without investigating the meaning of those terms. In Twitter, hashtags form the community driven convention for classifying Tweets created originally by Twitter users. From such communities, we can infer relations and high level categories for terms without any semantic analysis of those terms.

In this paper we describe an approach to dynamically discovering possible high-level categories and sub-categories, which result in possible taxonomies for different terms, by exploiting the wisdom of crowds captured from Twitter users. The main goal is to disambiguate a term, not by its semantic meaning, but through crowd sourced data from communities established in the micro-blogging sphere. We propose a method which utilizes collections of hashtag to discover terms related to a keyword and infer high-level categories for that keyword. The method builds upon possible taxonomies for terms based on collections of hashtags mentioned together in Twitter and based on the Open Directory Project (ODP) database.

The remainder of this paper is structured as follows. First we present related work, and then we describe the overall problem and approach. Afterwards, we present the results of applying the proposed approach on a sample of terms. Finally, we relate our work with other possible fields and draw conclusions for future work.

II. Related Work

User-generated content and the wisdom of crowds are terminologies which started to appear as web 2.0 technologies emerged. People publish and generate content individually which results in huge amounts of heterogeneous data stored on and transmitted over the web and especially in Social Network Services (SNSs). Research has shown that there is hidden knowledge behind these large datasets [11]. In media, for example, users' comments on news articles can give indications about sentiment views on political incidents [16]. Twitter posts have been used in research to predict election results in [2] based on sentiment opinions revealed by users' Tweets. In [1] Twitter messages were classified into positive or negative reviews of a product in respect to their authors' opinion.

978-1-4799-0048-0/13/$31.00 ©2013 IEEE

Wikipedia is an outstanding example of the success of user-generated content publishing. From Wikipedia articles and tags used to classify these articles, another successful project emerged called DBpedia. DBpedia derives ontologies from Wikipedia articles to discover relations between terms. Developers have been applying DBpedia knowledge in their applications which might be developed to serve different fields [7]. Researchers, trying to harness the power of large data generated by users on the web, have used both WordNet and DBpedia as references for enriching entities mentioned in users' posts. This has helped them in identifying connections between terms and creating categories and classification either for web documents containing these terms, or for users' posts in social networks, especially from the Twitter public streamline. The discovery of relations between entities recognized in the text of a Tweet is usually not based solely on other Tweets due to the noisy nature of its text.

The DBpedia knowledgebase has been used to enrich content of text, and the approach developed in [9] discovers relations between entities mentioned in user's Tweets by mapping entities to their corresponding DBpedia resources. In that approach, a pair was considered to be related if one entity is mentioned or referenced in one of the other's set of properties used in DBpedia to define a term such as full text description. Also, news articles published over the same time period were used to enrich Twitter posts and discover semantic relations in the entities recognized. In contrast, the approach developed in [3] is a graph based approach, where associations between entities are measured according to paths between two given entities in a graph based on DBpedia relations. The WordNet library, on the other hand, has been used in [12] to measure the semantic similarity between Tweets based on the relations between its discovered entities retrieved from the WordNet taxonomy.

The goal of finding similarities in users' Tweets can vary from detecting dynamic communities and trends to other user-centered goals, such as user modeling or recommendations of content or people. For example, the approach developed in [15] exploits user generated content in SNSs to detect dynamic communities. A collection of terms that co-occur together frequently over a period of time is an indication of a new trend emerging. Networks of associated terms are then represented as a graph where edges are weighted according to the frequency of them co-occurring together. The graph is then used as an input for a community detection algorithm to discover bursting events over a period of time. Furthermore, Tweets crawled from users' profiles have been investigated to determine whether a user model can be inferred from their Tweets. In [14], an approach called "Twopics" has been developed to categorize a Tweet based on the discovered entities in its text. A term is disambiguated according to its local context, where local context of a Tweet is formed by the entities that appear with the term in the same Tweet. Then, terms are mapped to the Wikipedia taxonomy to discover categories of topics that represent users' interests. Another approach has been applied in [13], where users of Twitter are ranked according their similarity to another user based on their interest. The goal from this approach was to suggest users for potential collaboration in solving innovation challenges.

In the work described above, what is considered an entity differs from one approach to another. In some, any capitalized non-stop word is considered to be an entity [14], whereas in other approaches Named Entity Recognition (NER) techniques have been applied to discover entities. In [6], Tweets are considered as pure text without any specific entities, and suitable hashtags are recommended for a user's Tweet in real time. The approach of recommendation proposed in [6] depends on finding messages from a set of crawled Tweets which are similar to the Tweets just entered by a user, and the hashtags existing within the similar Tweets are then extracted in order to be ranked according to their popularity in the Tweets retrieved. Top hashtags are then recommended for the user.

In our approach, we only consider hashtags as entities. Related Tweets are retrieved from Twitter search feed when searching for a specific hashtag and all the Tweets mentioning the same hashtag in its text are considered related Tweets. Finding related Tweets and extracting hashtags help in building a vocabulary of related terms which will be compared against the ODP database in order to discover high-level categories that overlap between the set of related terms discovered from the collection of hashtags extracted and the hierarchies which are built into the ODP database. This work is semi-supervised due to the spontaneous nature of Twitter data and is fully based on user generated content from Twitter and ODP.

III. Problem and Overall Approach

In this study, we use wisdom of the crowds that can be captured from Twitter users to discover related terms (topics) mentioned by people. Exploiting user generated content and integrating the content in order to find collections of frequently co-occurring terms allows us to discover related topics without looking into their semantic relationships.

The associations between terms are based on the numbers of co-occurrences of terms in the same collection of Tweets crawled. A collection of terms is considered related to a keyword if the terms tend to co-occur to a high degree with that keyword in the collection of Tweets. Keywords and terms in this work are considered to be the hashtag terms mentioned in Tweets.

The method proposed is divided into two steps: (a) crawling and retrieving collections of Tweets relating to a specific keyword and extracting related terms from the Tweets; (b) discovery of high-level categories that encompass most of these terms.

In the first step, we track Tweets from the public stream of Twitter and retrieve the ones containing a specific term (keyword hashtag) in its hashtags list. A collection of hashtags that co-occur with the keyword hashtag, in the same Tweet, are then extracted and the frequently occurring set of hashtags is considered to be the set of terms associated with that keyword. In the second step, the collection of related terms discovered in step one are mapped against the ODP database in order to infer categories and sub-categories where most of the terms overlap.

We thus classify each term in possible taxonomies dynamically.

A. Data Collection

For the first step, we track the public stream of Twitter using the Application Programming Interface (API) provided by Twitter (http://dev.twitter.com). The filter API allows us to track and retrieve public Tweets with specific query criteria. We filtered the Tweets and retrieved the ones that contain a particular term in their hashtag lists.

In the second step, in order to find a category for a set of terms, we used the Open Directory Project (ODP) database which is a freely available dataset by DMOZ (http://ww.dmoz.org). This directory is arguably the largest, most comprehensive human edited directory on the web. ODP is also known as DMOZ, an acronym for Directory Mozilla. It was founded by Netscape in the spirit of open source movement, and so its full dataset is available for download free of charge.

B. Detailed Approach

The specific steps followed in this method, for inferring possible taxonomies for a term (keyword), are listed below.

- Track and retrieve all Tweets containing the keyword in its hashtags list over a period of time.
- Pre-process the Tweets collected; remove the re-Tweets, remove punctuation, count Tweets collected.
- Extract hashtags from the collected Tweets, and keep track of the number of occurrences of the keyword for each hashtag. Due to the spontaneous nature of Tweets, and the systematic way for extracting them, human refinement is needed to find similar terms or abbreviations used to describe the same term. For example, C++ and cplusplus refer to the same term which is a programming language.
- Weight each hashtag, according to its occurrence frequency in the set of Tweets collected using the Inverse Document Frequency (IDF) measure. In Information Retrieval (IR) this weighting schema assumes that there are *N* documents in the collection, and that term *t* occurs in *ni* of them. Then the Inverse Document Frequency (IDF) for a term *idf*(*t*) is calculated using equation 1.

$$idf(t) = \log (N/ni) \qquad (1)$$

- The lower the value of *idf*, the more frequent the term appears in the collection. This weight schema is applied for weighting hashtags in this method where each Tweet is treated as a document.
- Split the collection of hashtags retrieved into two groups; frequent and non-frequent terms. We used the median value for distinct IDF values as a threshold in this process due to the large number of non-frequent hashtags in the collection.
- The final step is discovering possible high-level categories and sub categories that form taxonomies for the keyword term. In this step, associated terms (frequent hashtags) are reflected against the ODP database in order to retrieve all possible taxonomies for the terms. Then, finding the common categories in each level results in building the taxonomy for the keyword term.

IV. Experiments and Results

In this section, we present the results of several experiments conducted to validate the proposed method. Terms applied in these experiments were selected from tags used to classify articles posted in the Warwick Knowledge Centre website (http://www.warwick.ac.uk/knowledge). The centre is a gateway to the University of Warwick's world class expertise, research and learning, and by using such a controlled environment, we had access to a large number of articles containing relevant and timely content, and therefore had confidence on the accuracy of the tagging applied to the articles. A sample from the terms used in tagging articles in the Health and Medicine subject area was used to test the validity of the proposed method. When choosing this sample, terms were avoided which represent larger categories since the goal of this method is to expand knowledge about a specific subject.

From the collection of tags found ("Medicine", "Health", "Genetics", "Biology", "Obesity", "Surgery", "Smoking", and "Alzheimer"), we selected the more specific terms and retrieved Tweets having any of these terms in its hashtags list. The terms and the number of Tweets collected having each term, are shown in Table I.

Table I shows the number of Tweets retrieved after tracking the public stream of Twitter for 24 hours only. Only Tweets written in English and containing one of the terms in its hashtags list were retrieved. Re-Tweets were removed from the collection retrieved so that they do not affect the frequency of the hashtags calculated in the next step.

After collecting the sample using the Twitter API, hashtags were extracted from the Tweets with their frequency of appearing in the collection calculated. Table II details the number of hashtags extracted from each collection of Tweets and the numbers of popular hashtags.

TABLE I. Keywords and the numbers of Tweets retreived for each keyword

Keyword	No. Of Tweets
Genetics	250
Obesity	693
Smoking	663
Alzheimer	226

TABLE II. Numbers of hashtags co-occurring with each keyword

Keyword	No. of Distinct hashtags	No. of Popular hashtags
Genetics	251	9
Obesity	485	12
Smoking	1159	16
Alzheimer	153	4

The threshold used to split the collection of hashtags into popular and less popular ones is explained in the methodology. After weighting the hashtags according to their frequencies in the collection, the threshold is calculated for each collection of hashtags. Figure 1 demonstrates the weights of hashtags related to the keyword "Obesity", where the IDF values are represented on the x-axis, and y-axis represents the collection of hashtags. The dashed line represents the threshold which is the median value of distinct IDF values in that collection of hashtags.

In figure 1, the hashtags are represented by small dots. It is noticeable that a large number of hashtags have high IDF values, which indicates that these hashtags are less frequent than the ones with low IDF weights. The 12 popular hashtags, for that collection presented in Table II, are shown to the left of the dashed line.

In the final step, popular hashtags were mapped to the ODP database to find the common category between the co-occurring hashtags if it exists. The top category is the one that overlaps the maximum number of the popular hashtags. Table III illustrates the top categories resulted from mapping the popular hashtags for each keyword into the ODP database.

In Table III, we notice that not all popular hashtags are listed in the ODP database. That is due to the spontaneous nature of Twitter hashtags authored by users. For example, 16 popular hashtags occur with the keyword "Smoking", but only 6 of them are listed in ODP database. Terms such as "Ecigarette", "Quit", "High", "Weed", etc., are popular but are not listed in the ODP database.

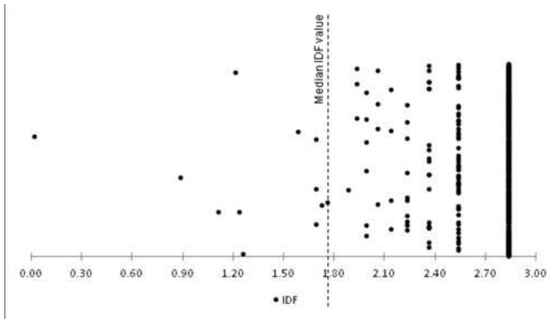

Fig. 1. Example of splitting "Obesity" related hashtags using median IDF value.

TABLE III. Results of reflecting the collection of Terms into ODP

Keyword	No. of Popular hashtags	No. of hashtags found in ODP	No. of common hashtags in Top Categories	Top Categories
Genetics	9	8	6	Science Society Reference
Obesity	12	6	6	Health
Smoking	16	6	4	Health Society
Alzheimer	4	2	2	Health

For each hashtag that is listed in the ODP directory, several possible directories that this term belongs to are retrieved from the ODP database. At this point, we have a set of directories pointing at set of terms. The next step is to find the mutual category for all these terms, and this is achieved by finding the common root between the sets of directories retrieved for each term or the root where most of these terms overlap. The top category is the root that overlaps most of the popular hashtags directories retrieved from the ODP database. For example, in the collection of hashtags related to "Obesity", 6 terms are found in the ODP database and the 6 terms have the same root in common, which is "Health". For the terms related to the "Genetics" keyword, 8 terms appeared in the ODP database and 6 of them have common roots "Science", "Society", and "Reference". In the "Genetics" and "Smoking" examples, we can see that there is no specific top category since all the top categories found in the root of directories have the same number of hashtags in common. This problem can be solved when more data are collected, which will change the frequencies for the popular hashtags.

Also, the full directory retrieved from the ODP database can be used to discover sub categories for larger categories, not only the root. In the case of "Alzheimer", for example, "Health" is the top category that is discovered, and the 2 popular hashtags were "Dementia" and "Alzheimer". If we look at the full directories retrieved from ODP when searching for these two terms, we see that both are descendents of "Conditions_and_Diseases" and "Neurological_Disorders" from the same root "Health". In this case, it is easy to find the exact taxonomy for the word "Alzheimer". In other cases such as "Genetics" we can discover a set of sub categories that are a possible fit for the world "Genetics" by finding the most frequent sub categories in the first level after the root and the second level and so on. Figure 2 illustrates how we can surround the word "Genetics" with subcategories.

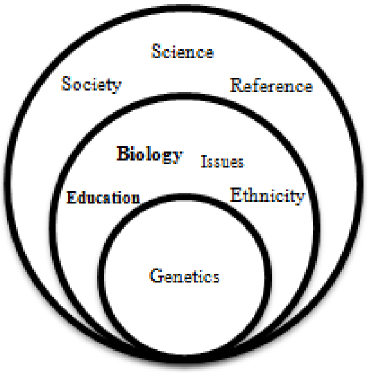

Fig. 2. Categories and sub-categories induced from this method for the term "Genetics"

The sub categories for the word "Genetics" in Figure 2 are inferred using the same technique for discovering the root. The most common terms which appear in the first level of directories after the root in the ODP database having one of the top categories as its root are considered the top sub categories. Looking at the sub-categories in figure 2, one can infer possible taxonomies for genetics, such as, ("Genetics", "Biology", "Science"), ("Genetics", "Ethnicity", "Society"), or ("Genetics", "Education", "Reference").

V. Conclusion and Future Work

In this paper we have presented a methodology for inferring dynamic taxonomies for a seed term by extracting hashtags from Tweets in Twitter search feed when searching for the seed term and from ODP databases. Our analysis of the taxonomy inferred for the terms tested in this study (we used "health and medicine" as the theme) revealed four high-level categories. These categories were consistent with the subject themes used to structure the site from which the seed terms were sourced, and this suggests that our proposed methodology can accurately generate taxonomy.

Exploiting the wisdom of crowds in this method to infer relations between terms will be beneficial for discovering hierarchies for terms in disciplines such as computer science where some terminologies do not have semantic meaning, for example, "Java" , "J2ME", "Python", etc. Furthermore, relations between such terms are hard to discover unless ontologies exist to summarize this knowledge, and these are not available in every discipline. Future work will investigate the potential to develop this method to utilize the vocabularies resulting from all possible dynamic taxonomies inferred. The application of this method in different areas such as adaptive education, where the dynamic taxonomies inferred can be used in authoring adaptive websites, will also form part of our future research.

References

[1] A. Go, R. Bhayani., and L. H. Huang, "Twitter sentiment classification using distant supervision," CS224N Project Report, Stanford, 2009.

[2] A. Tumasjan, T. O. Sprenger , P. G. Sandner , and I. M. Welpe , "Predicting elections with Twitter : What 140 characters reveal about political sentiment, " in In Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media , 2010, pp. 178-185.

[3] B. P. Nunes, R. Kawase, S. Dietze, D. Taibi, M. Casanova, and W. Nejdl , "Can entities be friends ?," in Proceedings of Web of Linked Entities (WOLE2012), Workshop at the 11th International Semantic Web Conference (ISWC2012), 2012.

[4] D. Niko, H. Hwang, Y. Lee, C. Kim, "Integrating user-generated content and spatial data into Web GIS for disaster history ," in Computers, Networks, Systems, and Industrial Engineering 2011, Studies in Computational Intelligence, 2011, pp. 245-255.

[5] E. T. Bradlow, "User-generated content : The 'voice of the customer' in the 21st century," in Marketing Intelligent Systems Using Soft Computing, Springer, 2010, pp. 27-29.

[6] E. Zangerle, W. Gassler, and G. Specht, "Recommending #-tags in Twitter," in Proceedings of the Workshop on Semantic Adaptive Social Web (SASWeb), 2011, pp. 67-78.

[7] G. Kobilarov, C. Bizer, S. Auer, and J. Lehmann, "DBpedia - A Linked Data Hub and data source for Web and enterprise applications," in Proceedings of the Developers Track of 18th International World Wide Web Conference WWW 2009, 2009, pp. 1-3.

[8] H. Ickler, and U. Baumol, "Adding value with collective intelligence – a reference framework for business models for user-generated content," Advances in Collective Intelligence 2011, Springer,2012, pp. 35-52.

[9] I. Celik, F. Abel, and G. Houben, "Learning semantic relationships between entities in Twitter," in Web Engineering, 2011, pp. 167-181.

[10] J. Vom Brocke, C. White, U. Walker, and C. Vom Brocke, "Making user-generated content communities work in higher education – The importance of setting incentives," in Changing Cultures in Higher Education, Springer, 2010, pp. 149-166.

[11] K. Uchimura, and A. Nadamoto, "Extracting hidden information based on comparing Web with UGC," Web Information Systems Engineering-WISE, 2010, pp. 365-377.

[12] L. Li, H. Xiao, and G. Xu, "Finding related micro-blogs based on WordNet," Database Systems for Advanced Applications, 2012, pp. 115-122.

[13] M. Stankovic, M. Rowe, and P. Laublet, "Finding Co-solvers on Twitter , with a Little Help from Linked Data," The Semantic Web: Research and Applications, 2012, pp. 39-55.

[14] M. Michelson, and S. A. Macskassy, "Discovering users' topics of interest on Twitter : A first look," in Proceedings of Fourth Workshop on Analytics for Noisy Unstructured Text Data, ACM, 2010, pp. 73-80.

[15] R. Cazabet, H. Takeda, M. Hamasaki, and F. Amblard, "Using dynamic community detection to identify trends in user-

generated content," Social Network Analysis and Mining, 2012, vol. 2, no. 4, pp. 361-371.

[16] S. Park, M. Ko, J. Kim, Y. Liu, and J. Song., "The politics of comments : Predicting political orientation of news stories with commenters' sentiment patterns," in Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW'11), 2011, pp. 113-122.