Exploiting Big Data for Enhanced Representations in Content-Based Recommender Systems

Fedelucio Narducci[2], Cataldo Musto[1], Giovanni Semeraro[1], Pasquale Lops[1], and Marco de Gemmis[1]

[1] Department of Computer Science
University of Bari Aldo Moro, Italy
name.surname@uniba.it

[2] Department of Information Science, Systems Theory, and Communication
University of Milano-Bicocca, Italy
narducci@disco.unimib.it

Abstract. The recent explosion of Big Data is offering new chances and challenges to all those platforms that provide personalized access to information sources, such as recommender systems and personalized search engines. In this context, social networks are gaining more and more interests since they represent a perfect source to trigger personalization tasks. Indeed, users naturally leave on these platforms a lot of data about their preferences, feelings, and friendships. Hence, those data are really valuable for addressing the cold start problem of recommender systems. On the other hand, since content shared on social networks is noisy and heterogeneous, information extracted must be hardly processed to build user profiles that can effectively mirror user interests and needs.

In this paper we investigated the effectiveness of external knowledge derived from Wikipedia in representing both documents and user profiles in a recommendation scenario. Specifically, we compared a classical keyword-based representation with two techniques that are able to map unstructured text with Wikipedia pages. The advantage of using this representation is that documents and user profiles become richer, more human-readable, less noisy, and potentially connected to the Linked Open Data (LOD) cloud. The goal of our preliminary experimental evaluation was twofolds: 1) to define the representation that best reflects user preferences; 2) to define the representation that provides the best predictive accuracy.

We implemented a news recommender for a preliminary evaluation of our model. We involved more than 50 Facebook and Twitter users and we demonstrated that the encyclopedic-based representation is an effective way for modeling both user profiles and documents.

Keywords: User Profiling, Social Network Analysis, Explicit Semantic Analysis, Personalization, Facebook, Twitter, Recommender System.

C. Huemer and P. Lops (Eds.): EC-Web 2013, LNBIP 152, pp. 182–193, 2013.

1 Introduction

Social networks have rapidly changed the interaction among people, thus becoming a real hub of information shared on the Web. A recent statistic reports that 91% of online adults use regularly social networks: every minute 100k tweets are sent and 684,478 pieces of content are shared on Facebook[1]. Even though the original aim of social networks was merely to allow friends to keep in touch, nowadays these platforms are becoming really valuable mines of information about user preferences, which can be exploited by personalization systems. On the other hand, content shared on social networks is noisy and heterogeneous, and must be deeply processed to extract information which mirrors effectively the preferences and interests of users. However, even though the gathering and representation of user interests play a crucial role, equally important is the representation of items to be recommended.

In this paper we investigated the effectiveness of a representation based on Wikipedia concepts (articles) both for user profiles and items in a recommendation scenario. Accordingly, we associated to each user profile a set of Wikipedia concepts most related with the user interests. The same process was performed on the items (i.e., news). We guess that this kind of representation brings in different advantages such as: producing more transparent and human-readable user profiles, removing noise, making profiles and items ready to be easily connected to the Linked Open Data (LOD) cloud.

We thus analyze two different aspects: first, we try to define the best model for representing user interests; second, we compare different groups of recommended news by representing content in different ways.

The rest of this paper is organized as follows. Section 2 analyzes the state of the art. Section 3 introduces the techniques adopted for obtaining a Wikipedia-based representation of content. Section 4 defines the representation of profiles and documents, and in Section 5 the recommendation model is introduced. Finally, experimental results are presented in Section 6, and in Section 7 the conclusion and the future work are summarized.

2 Related Work

In literature, several works try to model user profile by mining data extracted from social networks. In [7] the authors present a methodology for building multifaceted user models from raw Twitter data. Tweets are also exploited in [11] to model users in a news recommendation scenario. In the same domain, Abel et al. [1] model user interests in terms of entities, hashtags and topics a tweet refers to. Next, in [10] the authors propose a methodology for modeling profiles of user interests by extracting information by Facebook, Twitter, and LinkedIn as well. The most distinguishing aspect of our approach lies in the fact that we adopt a Wikipedia-based text representation which allows the construction of

[1] http://thesocialskinny.com/216-social-media-and-internet-statistics september-2012/

more transparent and human-readable user profiles, rather than using a simple keyword-based representation of the content extracted from the social networks. A strategy for linking user interests to Wikipedia is presented in [14], where the authors elicit user preferences by analyzing their personal folksonomy on delicio.us, then tags are connected to Wikipedia categories. Another strength that comes from the adoption of a Wikipedia-based representation of user interests is that each facet of the user profile can be easily linked to the LOD cloud by exploiting DBPedia[2], thus enabling a sort of reasoning on the information stored in a user model. Furthermore, a system that adopts a more understandable representation can lead towards a more transparent personalization process. For example, a recommender system that uses a human-understandable profile could easily explain the reason for a suggestion, and, as stated in [12], transparency is an essential feature of personalization tasks. Differently from the approach presented in [10], where a reasoning process leverages domain ontologies for inducing new implicit interests, our strategy for the generation of new topics exploits Explicit Semantic Analysis (ESA)[6].

Wikipedia-based document representations are adopted in different areas such as document similarity, information retrieval, and clustering. In [5] ESA is adopted for computing semantic relatedness between documents. Authors demonstrated that the proposed Wikipedia-based representation is more effective than a keyword-based one for that specific task. Similar results are also confirmed in [9,15] . A Wikipedia-based representation leverages ESA and outperforms a keyword-based document representation [2] also in an information retrieval scenario. Wikipedia is effectively exploited for cross-lingual and multilingual information retrieval, as well [13]. Finally, in [8] authors show that clustering performance significantly improves by enriching document representation with Wikipedia concepts and categories. To the best of our knowledge there are no work in literature that exploit a Wikipedia-based representation for addressing personalization tasks.

3 Wikipedia-Based Text Representation

Two different techniques were exploited for obtaining Wikipedia-based content representation: the anchor disambiguation algorithm implemented in Tag.me[3] and the Explicit Semantic Analysis (ESA) [6].

Tag.me - TAG.ME is an online tool developed by the University of Pisa (Italy) that implements an *anchor disambiguation* algorithm. It produces a Wikipedia-based representation of short text fragments, where the most relevant concepts occurring in the text are mapped to the Wikipedia articles they refer to, according to inter-relations between Wikipedia pages, as well as other heuristics. More details about the approach are provided in [4].

ESA - ESA is a vectorial representation of text, proposed by Gabrilovich and Markovitch [6], that uses Wikipedia as a space of concepts explicitly defined

[2] DBpedia is a RDF-based mapping of Wikipedia - `http://dbpedia.org`

[3] `http://tagme.di.unipi.it/`

and described by humans. The idea is that the meaning of a generic term (e.g. *London*) can be described by a list of concepts it refers to (e.g. the Wikipedia articles for: *London Eye, Big Ben, River Thames*). Formally, given the space of Wikipedia concepts (articles) $C = \{c_1, c_2, ..., c_n\}$, a term t_i can be represented by its *semantic interpretation vector* $v_i =< w_{i1}, w_{i2}, ..., w_{in} >$, where weight w_{ij} represents the strength of the association between t_i and c_j. Weights are obtained from a matrix T, called ESA-*matrix*, in which each of the n columns corresponds to a concept (Wikipedia article), and each row corresponds to a term of the Wikipedia vocabulary, i.e. the set of distinct terms in the corpus of all Wikipedia articles. Cell $T[i, j]$ contains w_{ij}, the TF-IDF value of term t_i in the article (concept) c_j. The semantic interpretation vector for a text fragment f (i.e. a sentence, a document, a tweet, a Facebook post) is obtained by computing the centroid of the semantic interpretation vectors associated with terms occurring in f.

The main difference between the two approaches is that ESA can generate *new* features related to the text to be indexed, while TAG.ME simply detects Wikipedia concepts that actually occur in the text. Hence, the former performs a *feature generation* process, while the latter performs a sort of *feature selection*.

4 Profile and Document Representation

We exploited TAG.ME and ESA in order to semantically annotate user profile and news. In the former case the input are the data about the user interests extracted from Facebook and Twitter, in the latter the input is a set of news titles coming from RSS feeds.

4.1 Wikipedia-Based Profile Representation

In the following, we describe a component called *Social Data Extractor* designed for this purpose. It is able to extract the following textual information about user activities on Facebook and Twitter:

- FACEBOOK: title and description of liked groups, title and description of attended events, title and description of liked pages, personal statuses, liked statuses, title and summary of shared links;
- TWITTER: personal tweets, tweets of followings, favorite tweets, direct messages.

For the sake of simplicity, all the aforementioned pieces of information will be identified hereafter by the expression *social items*. Three different kinds of profile representations are obtained by processing social items with the techniques described in the previous section. Examples of profiles, shown as tag clouds, are given in Figure 1.

Fig. 1. Examples of user profiles

Social Profile - This is the simplest representation since it is based merely on the keywords occurring in the social items collected for the user: only tokenzation and stopword elimination were applied, while the weight associated with each keyword is just its TF-IDF score. The social profile is the baseline for the other models.

Tag.me Profile - This representation leverages the algorithm implemented in TAG.ME to identify the Wikipedia concepts that occur in the social profile. Given a set of social items for user u, TAG.ME identifies those that can be mapped to Wikipedia concepts. All the titles of the identified Wikipedia concepts are included into the TAG.ME profile of u. The weight of each Wikipedia concept is the TF-IDF score of the keyword it refers to.

ESA Profile - This representation exploits the semantic interpretation vectors associated with keywords in the social items in order to identify *new* keywords which can be included in the profile. For each social item, the feature generation process is performed and the corresponding semantic interpretation vector is built (as described in Section 3 for text fragments). The 10 most relevant concepts, i.e. those with the highest weights in the semantic interpretation vector, are selected and the *titles* of the corresponding Wikipedia pages are included in the profile, together with their TF-IDF scores.

As an example, let's consider some statuses posted by a Facebook's user: *I'm in trepidation for my first riding lesson!*, *I'm really anxious for the soccer match :(* , *This summer I will flight by Ryanair to London!*, *Ryanair really cheapest company!*, *Ryanair lost my luggage :(*, *These summer holidays are really amazing!*, *Total relax during these holidays!*. The *Social Data Extractor* extracts and processes that information, by producing the profiles reported in Figure 1 (please consider that also other *social items* contribute to build those tag clouds). It emerges at-a-glance that the *social* profile is the richest one, since it also contains many non-relevant concepts, such as those referring to user moods (anxious, trepidation, etc.). On the other hand, the TAG.ME profile contains the terms that already occur into the *social* profile (horse, London, soccer, etc.), but their weights are higher since all the noise coming from non-relevant keyword has already been filtered out. Finally, in

the ESA profile there are some topics in some way related to the other profiles (riding horse, trip, Vienna[4]), but not *explicitly* mentioned in the Social profile. This is due to the fact that ESA enriches the basic representation with novel concepts associated with social items.

4.2 Wikipedia-Based Document Representation

Also for the documents, we compared a keyword-based representation with two Wikipedia-based models.

Keyword - This representation is only based on keywords. A *bag of words* is built for each news title. Tokenization, stemming and stopword elimination are performed on the text.

Tag.me - This representation is based on Wikipedia-concepts. The news title is the input to TAG.ME. Hence, TAG.ME identifies the Wikipedia concepts occurring in that text fragment.

Tag.me + ESA - This representation is obtained by combining TAG.ME and ESA results. The previously shown TAG.ME-based representation is enriched of Wikipedia concepts generated by ESA. Therefore, every news is represented by merging the Wikipedia concepts identified by TAG.ME and the Wikipedia concepts generated by ESA. The input to ESA is the news title and the 10-most related Wikipedia concepts are extracted from its semantic interpretation vector.

The motivation behind the combination of ESA with TAG.ME in a single profile is that sometimes ESA is not able to generate concepts for very short text fragment (several heuristics are applied in order to reduce the ESA-matrix dimension). Hence, we decided to have TAG.ME as basic representation and enrich it with the ESA concepts.

Since we need an unified representation both for documents and user profile, for each document representation we exploited the corresponding user profile built in the same way. Therefore, the *Social* profile is used to recommend news represented by the *Keyword* representation; the *Tag.me* profile is used to recommend news represented by the *Tag.me* model, and finally a profile obtained by merging the *Tag.me* and the *ESA* profile is used for recommending news adopting the *Tag.me + ESA* representation.

As an example, given the news title[5] "*At Facebook, Still the Undisputed Boss*". TAG.ME only identifies the Wikipedia page FACEBOOK; conversely the semantic interpretation vector generated by ESA contains the following Wikipedia concepts: FACEBOOK PLATFORM (the platform which enables third-party developers to integrate with the social network), SOCIAL GRAPH (term coined to describe "the global mapping of everybody and how they're related", on wich Facebook is based on), MARK ZUCKERBERG (the *undisputed boss* the news title refers to), DUSTIN MOSKOVITZ (co-founder of Facebook). This example confirms that ESA

[4] In Vienna is located the most world famous riding school.

[5] Extracted from the online version of The New York Times.

performs a feature generation process, while TAG.ME produces a sort of feature selection.

5 Learning Method

We implemented our recommender system as a text classifier. Hence, for each user we learned a classifier by exploiting the data extracted from social networks. The recommendation is thus a binary text classification task where the two classes are *like* and *dislike*. Subsequently, the learned classifier is used for deciding which items (i.e., news) are interesting (belonging to the class *like*) for a specific user. User feedback are exploited for updating the user profile and learning a new classifier. Probability as output is a really valuable feature in this context, since the recommender is able to perform a ranking of the suggested items (according to the probability to belong to the class *like*).

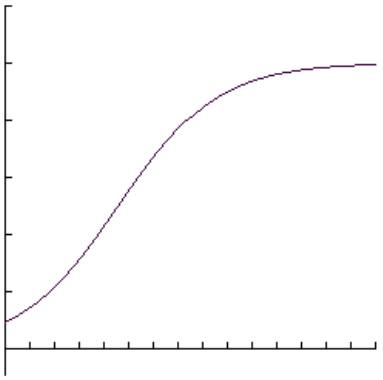

Fig. 2. Example of logistic function

We decided to use LOGISTIC REGRESSION as text classifier. LR belongs to the category of supervised learning methods. It is able to analyze data and recognize patterns and are used for classification and regression analysis. LR and its variants have been applied in several areas to solve classification problems. In [16] LR showed an accuracy comparable to SUPPORT VECTOR MACHINEs for several datasets with the advantage of yielding a probability model. The classification is performed by learning a logistic function on the training examples, that is represented by a sigmoid curve.

By analyzing Figure 2, on the x-axis we have the observed variable (e.g., the TF-IDF value), and on the y-axis we have the probability value (e.g., to belong to the class *like*). This is the simplest case in which we have only one feature, but it is easily extensible to more features.

After the model has been learned, new examples are mapped in the same previously built space and the correct class is chosen based on the value of the learned function.

In our experiments we use the LIBLINEAR library [3], an open-source library for large-scale linear classification (for datasets with a huge number of features and instances) that supports LR and SVMs with linear kernel.

6 Experimental Evaluation

We designed two different experimental sessions. The first one investigates the best model for representing user interests. In this session we are not interested in defining the user profile that achieve the best predictive accuracy, but we only focus our attention on the representation of user interests. The second session compares different groups of recommended news by representing content in different ways. We also evaluated the impact of the relevance feedback on the recommendation accuracy.

In order to avoid a cognitive overload of users, we invited a different group for each session. Each group was composed of 100 Italian Internet users. From the first group 51 users agree to participate to the first session: 36 gave us the consent to extract social items only from Facebook (71%), 4 only from Twitter (8%), 11 (21%) from both social networks. In the second session users were more unbalanced. 63 users of the second group accepted to participate: 62 Facebook users and only 1 Twitter user.

During the experiment users were driven by a wizard. Each session has been carried out for two weeks. Users were asked to login and to extract their data from their own Facebook and/or Twitter accounts (Figure 3). Next, in the first session, three user profiles were built according to the extracted data. Users were asked to rate the three profiles. In the second session, four groups of Italian news were proposed and users were asked to rate each group (Figure 4). After this step the user profiles were updated by exploiting the user feedback and four other news groups are proposed to rate. More details in the next sections. The two experiments took no more than five minutes per user. User votes were expressed by using a 5-star rating scale. The Wilcoxon Matched Pairs Test ($p < 0.05$) is used to test the significance of results (no assumption on the data distribution).

6.1 Session 1: Representation of Interests

The goal of the experiment was to identify which kind of user profile, among those discussed in Section 4.1, is the best representation for user interests. For each kind of profile, we defined the *transparency* as the overlap between actual user interests and keywords shown in the profile. For each user, the SOCIAL, ESA, and TAG.ME profiles were built and shown to her as tag clouds. Then, users were asked to answer to the following question, by using a 5-star rating scale:

1. How much the keywords in this profile reflect your personal interests and describe them in an accurate way?

For each representation, average rating, minimum and maximun ratings, and standard deviation are shown in Table 1. The representation obtained by TAG.ME

Fig. 3. Data Acquisition

describes the user interests in a better way than the other representations, as shown by the statistically significant differences among average ratings related to the transparency question. SOCIAL and ESA profiles obtained quite similar results (no statistically significant difference between them), while the ESA-based representation shows the highest standard deviation. Hence, it seems that this profile receives heterogeneous evaluations from users (also confirmed by the gap between MIN and MAX ratings). Indeed, ESA introduces new topics in the user profile, and this sort of *unexpectedness* is likely differently evaluated by the users.

Fig. 4. List of recommended news

6.2 Session 2: Representation of Documents

In this session we investigate how the document representation can affect the predictive accuracy of our recommender. Afterwards, we evaluated the impact of relevance feedback on the predictive accuracy. Also in this case we compare a keyword-based model with Wikipedia-based representations. Users were asked to evaluate four groups of recommendations, and for each group, five news were suggested. Each group of recommendations is generated by using one of the representation models defined in Section 4.2; the fourth group is the baseline of our experiment and is represented by random recommendations.

Results of this experimental session are reported in Table 2. The first outcome is that all the configurations have a statistically significant improvement with

Table 1. Results of Transparency and Serendipity Evaluation

Transparency				
Representation	Avg Rating	Min Rating	Max Rating	Stand. deviation
SOCIAL	1.33	0	3	0.65
TAG.ME	**3.88**	2	5	0.82
ESA	1.16	0	4	1.00

Table 2. Predictive Accuracy

1st cycle (without relevance feedback)				
Representation	Avg Rating	Min Rating	Max Rating	Stand. deviation
RANDOM	1.49	0	5	1.22
KEYWORD	1.89	0	5	1.47
TAG.ME	**2.86**	1	5	1.3
TAG.ME+ESA	2.59	0	5	1.37
2nd cycle (with relevance feedback)				
Representation	Avg Rating	Min Rating	Max Rating	Stand. deviation
RANDOM	1.49	0	5	1.12
KEYWORD	2.61	0	5	1.49
TAG.ME	**3.23**	1	5	1.35
TAG.ME+ESA	3.00	1	5	1.41

respect to RANDOM recommendations. For the first cycle, the highest average rating is achieved by using the TAG.ME representation. Differences between TAG.ME and both KEYWORD and TAG.ME+ESA representations are statistically significant. Hence, we can state that TAG.ME is an effective strategy for filtering out noise from the gathered content. The same results are confirmed in the second cycle (that exploits the user feedback), but in this case TAG.ME has a statistically significant difference only with the KEYWORD representation. TAG.ME+ESA shows a statistically significant difference with respect to KEYWORD, as well. Furthermore, also the difference between results of the first cycle and results of the second cycle is statistically significance. Finally, we can observe that there are not strong differences in terms of standard deviation and MIN/MAX rating among the different representations.

By summing up, even tough user feedback actually improve the predictive accuracy of the recommender, in a first stage where we have no explicit evidence from the user, the proposed Wikipedia-based representations are quite effective in modeling interests (gathered from social networks) and items of a recommender system.

7 Conclusions and Future Work

In this experimental evaluation we investigate different methods for representing user interests, and different methods for representing very short text (social items and news titles).

From a preliminary evaluation it emerged that users prefer a representation of their own interests expressed in terms of encyclopedic concepts with respect to simple keywords. The main outcome of the evaluation is that an encyclopedic-based representation of user interests that merges TAG.ME and ESA might lead to *unexpected* and transparent user profiles.

As regards the document representations, TAG.ME is an effective strategy for modeling items and user profiles. Also ESA significantly outperform the KEYWORD representation. Furthermore, the Wikipedia-based representations give the advantage of easy linking items and profiles to the LOD cloud.

In the future, we will investigate several weighing strategies in order to understand how the concepts coming from different sources can be merged. Furthermore, we want evaluate whether new topics introduced by ESA in the user profile can lead to serendipitous and unexpected recommendations. Finally, a comparison with other approaches based on the relationships encoded in the LOD cloud will be investigated.

Acknowledgments. This work fulfills the research objectives of the projects PON 02_00563_3470993 VINCENTE (A Virtual collective INtelligenCe ENvironment to develop sustainable Technology Entrepreneurship ecosystems) and PON 01_00850 ASK-Health (Advanced system for the interpretations and sharing of knowledge in health care) funded by the Italian Ministry of University and Research (MIUR).

References

1. Abel, F., Gao, Q., Houben, G.-J., Tao, K.: Analyzing user modeling on twitter for personalized news recommendations. In: Konstan, J.A., Conejo, R., Marzo, J.L., Oliver, N. (eds.) UMAP 2011. LNCS, vol. 6787, pp. 1–12. Springer, Heidelberg (2011)
2. Egozi, O., Markovitch, S., Gabrilovich, E.: Concept-based information retrieval using explicit semantic analysis. ACM Trans. Inf. Syst. 29(2), 8:1–8:34 (2011)
3. Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., Lin, C.-J.: LIBLINEAR: A library for large linear classification. Journal of Machine Learning Research 9, 1871–1874 (2008)
4. Ferragina, P., Scaiella, U.: Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, CIKM 2010, pp. 1625–1628. ACM, New York (2010)
5. Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: Proceedings of the 20th International Joint Conference on Artifical Intelligence, IJCAI 2007, pp. 1606–1611. Morgan Kaufmann Publishers Inc., San Francisco (2007)
6. Gabrilovich, E., Markovitch, S.: Wikipedia-based semantic interpretation for natural language processing. Journal of Artificial Intelligence Research (JAIR) 34, 443–498 (2009)
7. Hannon, J., McCarthy, K., O'Mahony, M.P., Smyth, B.: A multi-faceted user model for twitter. In: Masthoff, J., Mobasher, B., Desmarais, M.C., Nkambou, R. (eds.) UMAP 2012. LNCS, vol. 7379, pp. 303–309. Springer, Heidelberg (2012)

8. Hu, X., Zhang, X., Lu, C., Park, E.K., Zhou, X.: Exploiting wikipedia as external knowledge for document clustering. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2009, pp. 389–396. ACM, New York (2009)
9. Huang, L., Milne, D., Frank, E., Witten, I.H.: Learning a concept-based document similarity measure. J. Am. Soc. Inf. Sci. Technol. 63(8), 1593–1608 (2012)
10. Ma, Y., Zeng, Y., Ren, X., Zhong, N.: User interests modeling based on multi-source personal information fusion and semantic reasoning. In: Zhong, N., Callaghan, V., Ghorbani, A.A., Hu, B. (eds.) AMT 2011. LNCS, vol. 6890, pp. 195–205. Springer, Heidelberg (2011)
11. Phelan, O., McCarthy, K., Bennett, M., Smyth, B.: Terms of a feather: Content-based news recommendation and discovery using twitter. In: Clough, P., Foley, C., Gurrin, C., Jones, G.J.F., Kraaij, W., Lee, H., Mudoch, V. (eds.) ECIR 2011. LNCS, vol. 6611, pp. 448–459. Springer, Heidelberg (2011)
12. Sinha, R., Swearingen, K.: The role of transparency in recommender systems. In: CHI 2002: CHI 2002 Extended Abstracts on Human Factors in Computing Systems, pp. 830–831. ACM, New York (2002)
13. Sorg, P., Cimiano, P.: Exploiting wikipedia for cross-lingual and multilingual information retrieval. Data Knowl. Eng. 74, 26–45 (2012)
14. Szomszor, M., Alani, H., Cantador, I., O'Hara, K., Shadbolt, N.R.: Semantic modelling of user interests based on cross-folksonomy analysis. In: Sheth, A.P., Staab, S., Dean, M., Paolucci, M., Maynard, D., Finin, T., Thirunarayan, K. (eds.) ISWC 2008. LNCS, vol. 5318, pp. 632–648. Springer, Heidelberg (2008)
15. Yeh, E., Ramage, D., Manning, C.D., Agirre, E., Soroa, A.: Wikiwalk: random walks on wikipedia for semantic relatedness. In: Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing, TextGraphs-4, Stroudsburg, PA, USA, pp. 41–49. Association for Computational Linguistics (2009)
16. Zhang, T., Oles, F.J.: Text categorization based on regularized linear classification methods. Information Retrieval 4, 5–31 (2000)