A Semantics-Based Approach for Business Categorization on Social Networking Sites

Atia Memon, Christian Zinke, Kyrill Meyer

A Semantics-based Approach for Business Categorization on Social Networking Sites

Atia Bano Memon[1], Christian Zinke[2] and Kyrill Meyer[2]

[1]Department of Computer Science at the University of Leipzig, Leipzig, Germany
[2]Institute of Applied Informatics at the University of Leipzig, Leipzig, Germany

{memon, zinke, meyer}@informatik.uni-leipzig.de

Abstract. As the number and adoption of social networking sites (SNSs) supporting business representation in the form of business pages continues to escalate, more scalable and robust mechanisms for integrating data from different networks in order to serve the special purposes need to be envisaged. An important concern of such SNS data integration is the platform dependencies that different networks impose in collecting, organizing, and presenting the business information hosted on their servers. In this esteem, this paper deals with overcoming the challenge of different business categorization schemes being varyingly used by the existing SNSs. In doing so, we present a content-oriented approach for determining the business category on the basis of a semantic analysis of the textual information available in the business profile. The approach has been operationalized in the CoDiT (Company Discovery Tool) which is a web-based tool to facilitate the integrated business page search over multiple SNSs.

Keywords. Business page integration, SNS integration, Semantics-based category tagging, Business categorization

1 Introduction

Social networking sites (SNSs), a relatively new mechanism of human interaction, has gained a rapid popularity among a broader part of the population, many of whom have integrated these sites into their daily practices [1]. SNSs are the web 2.0 based applications that facilitate an online platform for the people to build social relations with other people and thus offer a new method of communicating, accelerating group formation, and escalating group scope and influence [2]. Today's SNSs originated in 1997 with the launch of SixDegrees.com that allowed the users to create profiles, list their friends, and add friends-of-friends to their own friend lists [1]. Since that time, the trend has gained such a rapid momentum that currently there exist several networks facilitating not only the interpersonal interaction, but also the online representation of all types of business organizations in the form of business pages (aka company pages). Through the business pages, the SNSs offer many opportunities

for the businesses that were previously not available or very difficult to acquire for many of them [3].

Increasingly, there exist a number of SNSs that offer the creation and hosting of business pages. Since these sites vary as to their scope as well as functionality [4], they cater a different type of audience. For example, whereas Facebook has a great appeal to the masses, LinkedIn is more focused on professionals [5]. Accordingly, the organizations are variously represented on different platforms in order to reach the audience that they target as Stelzner [6] has reported that the LinkedIn is most preferable platform among Business-to-Business (B2B) marketers while the Facebook is the most popular platform among Business-to-Customer (B2C) marketers. Since all of these platforms work in isolation from each other, useful information remains dispersed and confined to specific platform boundaries which in turn hampers the business discovery and information retrieval without actually getting on a specific platform [7]. Recently, however, the leading SNSs have started exposing their data for third-party applications through their Application Programming Interfaces (APIs). The social networking APIs facilitate the third party developers to fetch, aggregate, and transform SNS user data according to user's specific interests [8]. Accordingly, they enable the integration of business information available on different SNSs in order to serve special purposes and offer a data rich and seamless user experience.

While by leveraging the potential of SNS APIs, the platform boundaries of SNSs can be significantly dissolved, the issue of platform dependencies of the business information becomes more considerable. The existing SNSs support business representation in their own platform dependent ways and thus collect different pieces and types of business related information, impose different structure in information collection and presentation, employ varying search procedures, and offer different search management and page interaction functionalities [7]. A key challenge in this regard comes from the varying business categorization schemes being used by the existing SNSs. Specifically, the application developers have to envisage a mechanism for mapping the platform dependent categorization schemes of SNSs to a uniform scheme in order to support the efficient clustering and filtering of businesses in integrated result sets and thereby offer a seamless data integration. In this esteem, this paper presents a content-oriented approach for determining the business category on the basis of semantic analysis of the textual information available in the business profile. The approach has been operationalized in the CoDiT (Company Discovery Tool) which is a web-based tool for facilitating the integrated business page search over Facebook and LinkedIn platforms (cf. [7] for a detailed description of the tool). The methodology presented herein, besides addressing the issue of differing business categorization schemes of SNSs for the SNS business data integration, can also be applied by the existing SNSs themselves in order to uniformly tag the businesses represented on their servers.

2 Methodological Remarks

The work discussed in this paper is oriented around Design Science Research (DSR) approach which combining the disciplines of engineering and 'the science of the

artificial' [9] seeks to solve hitherto unresolved practical and theoretical problems by creating new and innovative artifacts [10]. The creation of DSR artifacts is initiated either to solve an identified problem, achieve an objective, implement a required functionality, or fulfill a client's requirement [11]. For the development of successful artifacts, various DSR methods have been suggested in the academic literature such as 'Build - evaluate - theorize – justify' [12], 'Problem identification - intervention - evaluation – reflection or learning [13]', 'Construct - develop - analyze - build - observe [14]', 'Awareness of problem - suggestion - development - evaluation – conclusion [15]', 'Develop or build – justify or evaluate [10]', and 'Problem identification and motivation - objectives of a solution - design and development - demonstration - evaluation – communication [11]'. The work described herein is aimed at solving a problem, and is undertaken in four steps drawn upon Takeda et al. [15] and Peffers et al. [11]: investigation and motivation of the problem and related challenges (section 3), designing a suitable solution of the identified problem (section 4), implementation of the proposed solution in a suitable context to demonstrate its usability (section 5), and evaluation of proposed solution to test its validity and performance (section 6).

3 Problem Description

The existing SNSs employ different structure for categorizing the business pages hosted on their servers. An examination of the four leading SNSs that facilitate the business page paradigm - Facebook[1], LinkedIn[2], Google+[3], and Xing[4] – indicates that the existing SNSs are adopting a structured approach for categorizing the business pages, i.e. they offer a predefined list of business categories and users are asked to select any of them to tag in their business profile. The Facebook, LinkedIn, and Google+ platforms employ a linear categorization scheme with more than 100 categories each. On the Facebook platform a business can be tagged with a maximum of 3 categories, the LinkedIn platform allows to select only one category (referred to as industry), and the Google+ platform allows to select any number of categories designating one as a primary category and others as additional categories. On the contrary, the Xing platform employs a hierarchical categorization scheme of 23 main categories with additional 121 sub-categories whereby a business can be tagged with a single parent category and one more subcategory relative to the parent category.[5]

Accordingly, while integrating business information from multiple SNSs the developers are faced with the question that 'how the business category information coming from different SNSs can be integrated into a uniform categorization scheme in order to support the efficient clustering and filtering of businesses in the integrated result sets?'. One obvious approach of addressing this is to manually map the

[1] https://www.facebook.com/
[2] https://www.linkedin.com/
[3] https://plus.google.com/
[4] https://www.xing.com/
[5] As on March 19, 2017

respective business categories of different platforms to a uniform category system. However, such an approach tends to be very tedious and time consuming as the developers have to recognize the categorization schemes of all the platforms that are intended to be connected and develop a mapping module for each platform individually. Moreover, the manual mapping approach brings the following three challenges:

1. The existing category schemes of SNSs include some very general categories such as 'Company', 'Organization', and 'Local business'. When the categories are mapped abstractly, it becomes impossible to categorize the businesses tagged with such broad categories into more meaningful and relevant categories.
2. When the business pages from different SNSs are integrated in order to address a specific task of a particular business domain, with the manual mapping it becomes challenging to further categorize the high-level (abstract) categories tagged to a business into more specific low-level (concrete) sub-categories in order to facilitate the drilling down to certain sub-domains.
3. When the category mapping is done explicitly, any future changes in the underlying categorization schemes of the integrated platforms are not captured by the application. Thus, the application becomes incapable to aggregate all the available business category information in an efficient and effective manner.

4 Semantics-based Business Categorization Approach

In this section, we propose a semantics-based approach for uniformly categorizing the business pages retrieved from different SNSs. According to the proposed approach, given a predefined business category list designating the uniform category scheme in which the business category information is to be represented (hereafter referred to as category list), relatedness of a business to the appropriate categories is determined on the basis of the semantic analysis of the concatenated textual information (hereafter referred to as business description) available in various free text fields of business profile such as 'about', 'general information', 'description', 'overview', 'products', 'specialties', and the like. As the category text usually consists of few words with no or minimal interdependence and all the words are equally important in determining category appropriateness, business categorization is considered as a classification task whereby one-to-one similarity decision is made. Accordingly, iterating through the given category list, the appropriateness of a particular category is determined in three steps: 1) preprocessing of category tag, 2) expansion of category tag, and 3) similarity measurement of the expanded category tag and the business description (cf. Fig. 1).

1. **Preprocessing of category tag:** During the first step, the category tag is initially normalized by converting it into lower case and removing any discrepancies present therein such as trailing whitespaces, or multiple spaces between the words. Subsequently, the normalized category tag is tokenized into individual words. For example, if a given normalized category tag is 'computer graphics', the tokenization returns an array (say T) of two words such that $T = \{T_1, T_2\} = \{$'computer', 'graphics'$\}$. Finally, the tokenized array of category tag is parsed to discard any stop words present therein, for example 'and', and 'of'.

2. **Expansion of category tag:** During the second step, the token list generated in step 1 is semantically and morphologically expanded. The semantical expansion corresponds to expanding the given token with the words similar to it in meaning (synonyms) in order to overcome the vocabulary mismatch problem. The semantical equivalents of a given word can be generated by different measures referring to either a suitable corpus or a standardized thesaurus. The morphological expansion corresponds to expanding the given token with its different lexical forms (singular/plural of nouns, different tenses of the verbs) in order to address the variances in the structure of sentences available in the business description. Accordingly, taking each token in the category tag, one at a time (say T_i), its semantically similar words (S_1 to S_n) are generated, and subsequently all morphological forms of the token itself (T_iM_1 to T_iM_n) and its semantical equivalents ($T_iS_1M_1$ to $T_iS_nM_n$) are generated. This step results into a bag of words (T) for the token T_i such that $T = \{T_i , T_iM_1 , \dots , T_iM_n , T_iS_1 , \dots . , T_iS_n , T_iS_1M_1 , \dots , T_iS_1M_n , \dots\dots\dots . , T_iS_nM_1 , \dots , T_iS_nM_n\}$.
3. **Measuring similarity:** During the third step, the inclusion of the T bag of words is determined in the business description through any appropriate text matching algorithm such that the similarity is positive if any of the words in T is present within the business description text. This process is iterated for each token of the category tag generated in step 1. The category is tagged to the business only if each token in any of its semantical or morphological form is present in the business description.

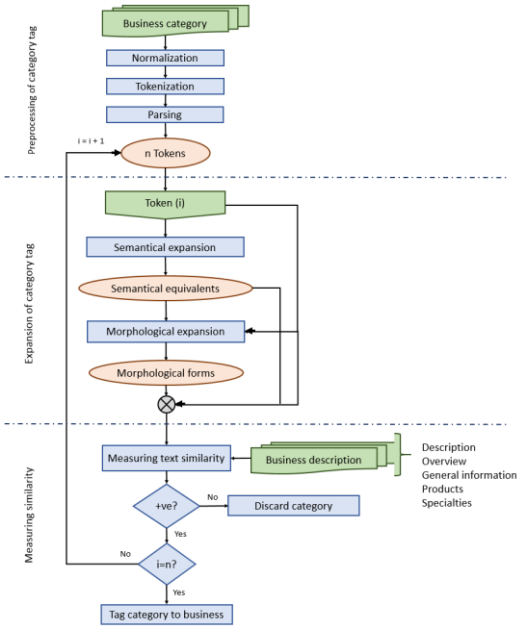

Fig. 1. Semantics-based business categorization approach

5 Implementation

The semantics-based business categorization approach given in section 4 has been implemented in a web-based tool – Company Discovery Tool (CoDiT) - that facilitates an integrated business page search over Facebook and LinkedIn platform together with business specific functionalities in order to support the business partner discovery for open innovation alliances. The key features of the tool include different types of search (keyword search with additional field restriction, advanced search in terms of business category, offered services, and location, and navigational search with respect to business category, offered services, and location), an integrated geographic map, faceted filtering in terms of business-specific attributes, search management functionalities such as bookmarking and networking the business pages, direct page interaction (liking/following, and reading/writing the feed), and uniform representation of the business information retrieved from Facebook and LinkedIn platforms (cf. [7] for more details).

In the CoDiT system, the proposed approach is applied to identify the business category as well as the services offered by a certain business. The business pages retrieved from the Facebook and LinkedIn are mapped to a uniform linear category scheme of 168 categories that includes the general category tags (e.g. company, organization), high-level business domains (e.g. computers, engineering, technology), and low-level specific categories (e.g. computer hardware, computer software, computer graphics). The CoDiT system implements the proposed approach as follows:

1. **Semantical expansion:** For generating the semantically similar words of the given tokens in the category tags, the CoDiT system employs the corpus-based web services available from Leipzig Corpora Collection[6] (LCC) using its REST API[7] (for a description of REST architecture, cf. [16]). LCC is the collection of a large corpus of freely available electronic media consisting of the web, newspapers, and Wikipedia. The resources are collected by employing different data collection methods. The resulting crawled resources are processed, and for each language, a full form dictionary with frequency information for each word is calculated. In addition, the significant co-occurrence statistics (the words that co-occur significantly often with a given word) are precomputed and two kinds of co-occurrence data are stored: words occurring together in sentences, and words found as immediate (left or right) neighbors [17, 18]. The LCC collects corpora for more than 200 languages and for each language different corpora releases are available. For the implementation of the CoDiT system, the English language corpus based on Wikipedia with 1M sentences (eng_wikipedia_2012_1M) is referenced. The 'similarity-service' of the LCC returns the words related to the given word on the basis of the similarity of their significant sentence and neighbor co-occurrences in the reference corpus.
2. **Morphological expansion:** For retrieving the different lexical forms of given tokens in the category tags, the CoDiT system leverages the PHP based

[6] http://corpora.uni-leipzig.de/
[7] http://wortschatzwebservices.informatik.uni-leipzig.de/ws/swagger-ui.html

phpMorphy library[8]. The phpMorphy is a morphological analysis library that provides the dictionary-based morphological services for the English, Russian, German, Ukrainian, and Estonian languages. For every word, phpMorphy provides three types of information: a base form of the word (lemma), all morphological forms of the word, and grammatical information of the word (part of speech, case, etc.).

3. **Similarity measurement:** For computing the similarity between tokens of the category tag and the business description text, the CoDiT system applies the cosine similarity measure which is the most common measure used for determining text similarity [19] and is the baseline for most of the similarity studies [20]. The basic idea behind cosine similarity is to transform each text string into a vector in some high dimensional space such that similar strings are close to each other. The cosine of the angle between two vectors is a measure of how similar they are, which in turn, is a measure of the similarity of given texts. Accordingly, at first, a high dimensional space V is created where each term in text T1 (the token, its similar terms, and morphological forms of the token and its similar terms), and text T2 (the business description) defines an independent dimension, such that $V = (T1_1, \ldots, T1_n) + (T2_1, \ldots, T2_n)$. Then, the texts (T1 and T2) are transformed into their respective binary vectors (V1 and V2) in this high dimensional space. Operating in the positive quadrant of the Euclidean space (i.e. no term is assigned a negative value), vectors represent the presence and absence of a particular term in each text string by a non-negative value (1 or 0) along the dimension corresponding to the term. Finally, the cosine of the angle between V1 and V2 is computed which is identical to their normalized inner product, such that

$$\text{Sim}_{\cos}(T1, T2) = \frac{V1 \,.\, V2}{\sqrt{V1^2}\,\sqrt{V2^2}}$$

Since, non-negative values are used in the vectors, the cosine measure returns positive numerical value in the range of 0 (for the orthogonal vectors) and 1 (for the identical vectors).

6 Evaluation

We have tested the validity of proposed approach on a data set of business descriptions of 10 randomly selected businesses represented on Facebook platform. In doing so, at first, we manually analyzed and annotated each business description with relevant categories from the predefined list of 168 categories. Subsequently, we compared the results of proposed approach (as implemented in CoDiT) with this manual standard. As shown in Figure 2, the proposed approach has an average recall of 0.85 (min. 0.67, max. 1), precision of 0.75 (min. 0.6, max. 1), and F-measure of

[8] http://phpmorphy.sourceforge.net/dokuwiki/

0.79 (min. 0.67, max. 0.91). Therefore, we are convinced that the proposed approach is adequate to achieve its set objective of semi-automatic business categorization on SNSs.

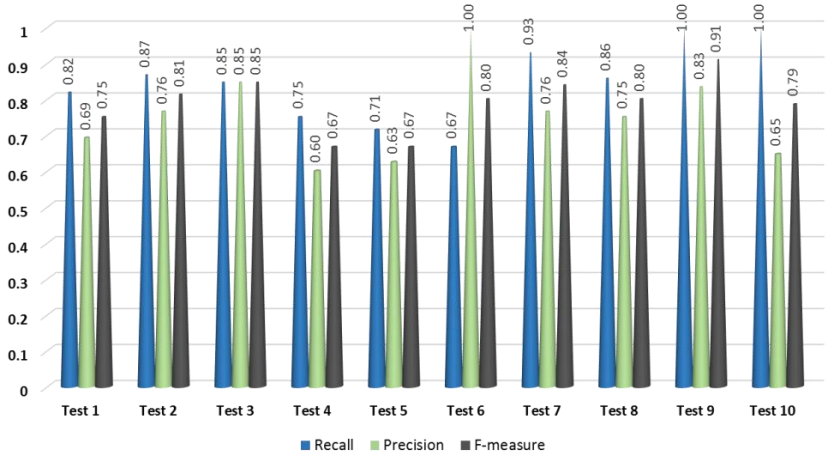

Fig. 2. Evaluation results of proposed approach

A screenshot of the CoDiT system showing the official Facebook page of IBM company[9] is given in Figure 3. As illustrated, the Facebook platform returns only one category for IBM, i.e. 'Company' (cf. category list in Figure 2)[10] which does not sufficiently help to understand what the given company deals with.

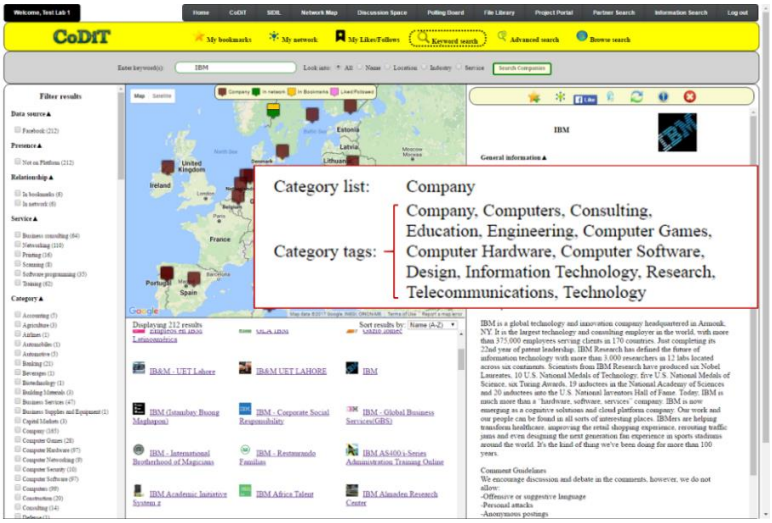

Fig. 3. Illustration of semantics-based business categorization approach

[9] https://www.facebook.com/IBM/
[10] As retrieved on March 22, 2017

The CoDiT system through the proposed approach has been able to determine several relevant categories for the company including the 'Company' (the general category), 'Computers', 'Consulting', 'Education', 'Research', 'Engineering', 'Technology', 'Telecommunications', 'Information technology', 'Design' (broad business domains), and 'Computer games', 'Computer hardware', and 'Computer software' (low-level categories) that substantially reveal the nature of the business and its offerings.

It is important to note here that while some of the given high-level categories such as 'Education' and 'Research' tend to be ambiguous and thereby need more drilling down to understand the area of education and research that the company operates in, the combination of such high-level categories with other high or low level categories implicitly declares the important relations such as 'Education' and 'Research' combined with 'Computers', 'Technology', and 'Engineering' indicates that the company would be dealing with the education and research in the fields of computers and information technology.

7 Conclusion

In this paper, we have addressed the issue of different business categorization schemes being varyingly used by the existing SNSs. In doing so, we have proposed a content-oriented approach for determining the relevant business categories on the basis of semantical and morphological analysis of the textual business information available in the business profile. The applicability and usability of the proposed approach is demonstrated within a web-based tool that supports the integrated business page search over multiple SNSs. The accuracy of the proposed approach depends on three factors: 1) the amount of information available in the business profile (business description), 2) the quality and depth of the corpus/thesaurus, and the morphological analyzer referenced to generate the synonyms and morphological forms respectively, and 3) the comprehensibility of the given business category list. However, given that these three components are appropriate, the approach is found to be adequate in attaining the task of semi-automatic characterization of businesses into the meaningful and appropriate categories.

References

1. Boyd, D.M., Ellison, N.B.: Social Network Sites: Definition, History, and Scholarship. *Journal of Computer-Mediated Communication* (2007). doi: 10.1111/j.1083-6101.2007.00393.x
2. Lin, K.-Y., Lu, H.-P.: Why people use social networking sites: An empirical study integrating network externalities and motivation theory. *Computers in Human Behavior* 27(3), 1152–1161 (2011)
3. Jefferson III, Carl E, Traughber, S.: Social Media in Business. *How Social Media Can Help Small Businesses and Non-Profit Organizations*, 2–3 (2012)

4. Kietzmann, J.H., Hermkens, K., McCarthy, I.P., Silvestre, B.S.: Social media? Get serious! Understanding the functional building blocks of social media. *Business horizons* 54(3), 241–251 (2011)
5. Papacharissi, Z.: The virtual geographies of social networks: a comparative analysis of Facebook, LinkedIn and ASmallWorld. *New media & society* 11(1-2), 199–220 (2009)
6. Stelzner, M.A.: 2014 social media marketing industry report: how marketers are using social media to grow their businesses. *social media examiner* (2014)
7. Memon, A.B., Meyer, K.: CoDiT. An Integrated Business Partner Discovery Tool Over SNSs. In: *Working Conference on Virtual Enterprises*, pp. 631–638 (2015)
8. Felt, A., Evans, D.: Privacy protection for social networking apis. *2008 Web 2.0 Security and Privacy* (W2SP'08) (2008)
9. Simon, H.A.: The sciences of the artificial, vol. 136. *MIT press* (1996)
10. Hevner, A.R., March, S.T., Park, J., Ram, S.: Design science in information systems research. *MIS quarterly* 28(1), 75–105 (2004)
11. Peffers, K., Tuunanen, T., Rothenberger, M.A., Chatterjee, S.: A design science research methodology for information systems research. *Journal of management information systems* 24(3), 45–77 (2007)
12. March, S.T., Smith, G.F.: Design and natural science research on information technology. *Decision support systems* 15(4), 251–266 (1995)
13. Cole, R., Purao, S., Rossi, M., Sein, M.: Being proactive: where action research meets design research. *ICIS 2005 Proceedings*, 27 (2005)
14. Nunamaker Jr, Jay F, Chen, M.: Systems development in information systems research. In: *System Sciences, 1990., Proceedings of the Twenty-Third Annual Hawaii International Conference*, pp. 631–640 (1990)
15. Takeda, H., Veerkamp, P., Yoshikawa, H.: Modeling design process. *AI magazine* 11(4), 37 (1990)
16. Feng, X., Shen, J., Fan, Y.: REST: An alternative to RPC for Web services architecture. In: *Future Information Networks, 2009. ICFIN 2009. First International Conference*, pp. 7–10 (2009)
17. Biemann, C., Heyer, G., Quasthoff, U., Richter, M.: The Leipzig Corpora Collection-monolingual corpora of standard size. *Proceedings of Corpus Linguistic* (2007)
18. Richter, M., Quasthoff, U., Hallsteinsdóttir, E., Biemann, C.: Exploiting the leipzig corpora collection. *Proceesings of the IS-LTC* (2006)
19. Huang, A.: Similarity measures for text document clustering. In: *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008)*, pp. 49–56 (2008)
20. Shrestha, P.: Corpus-based methods for short text similarity. In: *Rencontre des Étudiants Chercheurs en Informatique pour le Traitement automatique des Langues,* p. 297 (2011)