Extracting People's Hobby and Interest Information from Social Media Content

Thomas Forss, Shuhua Liu, Kaj-Mikael Bjork

Extracting People's Hobby and Interest Information from Social Media Content

Thomas Forss, Shuhua Liu and Kaj-Mikael Björk

Department of Business Administration and Analytics
Arcada University of Applied Sciences
Jan-Magnus Janssonin aukio 1 00560 Helsinki Finland
{thomas.forss, shuhua.liu, kaj-mikael.bjork}@arcada.fi

Abstract. In this study we investigate how to analyze people's social media profiles to extract hobby and interest information. We developed a baseline system that applies heuristic rules and TF-IDF term weighting method in determining the most representative terms indicating hobbies and interests. A pilot test was done to collect feedback from users concerning the perceived usefulness of the extracted tags. The baseline system was then extended to include new functionality to help set limits on the scope of relevant content, extract Named Entities, use of predefined dictionaries to identify even low-scoring hobbies and interests, and use of machine translation to handle content in multiple languages.

Keywords: social media analysis, term extraction, hobby and interest, Facebook profile, Named Entity extraction

1 Introduction

Since the beginning of the 21st century more and more users use some form of social media every day. Popular social media domains include blogs, web forums, photo and video sharing sites as well as social networking sites [1]. Popular social media networking sites that focus on sharing information with friends include but are not limited to Facebook, Twitter, LinkedIn, Google+, and Diaspora. As of January 2014 Facebook has over 1.3 Billion users spread all over the world while Twitter has about 645 Million users [2].

In parallel with the huge explosion of the usage of social media, we have an Analytics movement that strive to create competitive advantage and added value based on analyzing the huge amount of both structured and unstructured data. In Delen and Demirkan [3], the future of data, information and Analytics as a service is predicted. We will mostly probably see a significant number of new services emerging to help us analyze social media content.

Numerous research on social media analysis have been made, however there is much still to explore. Social media content analysis has been reported in several studies on shorter information snippets such as the 'tweets'. Shamma et al [4] built a tool, Statler that looks for trending topics, level of interest, and geo-locations of tweets. Zhao et al [5] extracts keywords and organizes the keywords according to topics learned from Twitter through a context-sensitive topical Page Ranking method. Yang et al [6] studied automatic summarization of Twitter tweets through topic modeling and event detection. Much less research has been done on analyzing and/or summarizing a more complete social media presence of people, i.e. user profiles, in sites like Facebook, LinkedIn, Diaspora, and Google+.

Text summarization and keyword extraction was originally done on structured documents that have paragraphs, sentences and correct grammar. In social media the format the text is saved in depends upon how the site or platform in question stores the data and how the user chooses to write. The challenge in analyzing a person's social media profile lies in that not all the available information is relevant to the user who is subject to the analysis, and not all the content is created by the one user. Additionally, the content can be fragmented and neither structured like normal text nor written in a grammatical way. Further, multilingual users tend to change languages between posts and can also sometimes write posts mixing several languages.

Clark & Araki [7] identifies 8 different grammatical and structural problems that can make extraction and/or summarization harder in social media content. Out of these 8 problems the following ones are relevant for this study: non-dictionary slang, punctuation omission/errors, intentional misspellings, and abbreviations. Bertoldi et al [8] extended a statistical machine translation tool with the capabilities to adjust for misspellings, combining the approach Clark & Araki [7] propose with machine translation could provide a similar result.

In this paper we present our work on analyzing social media profiles to extract users' hobby and interest information from Facebook content. We developed a baseline system that applies heuristic rules and TF-IDF term weighting method in determining the most representative terms indicating hobbies and interests. A pilot test was done to collect feedback from users concerning the perceived usefulness of the extracted tags. The baseline system was then extended to include functionalities that help limit the scope of relevant data, Named Entity recognition, and predefined dictionaries containing hobbies and interests, and possibilities to handle multiple languages.

The extensions of the system were added after the pilot test. As such the extensions while implemented have not been tested except for by the authors. We plan on running a second test with the extended system in the near future.

2 A baseline system and pilot test

2.1 A baseline system

To the best of our knowledge there are no systems available that extract or summarize content from a Facebook profile. So we started with building a baseline system. Our approach can be described as a three-step model shown in Figure 1. First we retrieve content and group it. Then we pre-process the content and lastly we extract hobbies and interests.

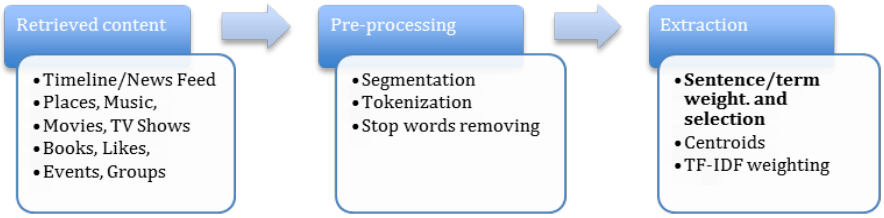

Figure 1 Determining the most representative terms for hobbies and interests in the baseline system.

Keyword/key phrase extraction help select a small set of words or phrases that are central to the information in the text, which in our case is the sum total of one person's activity on a social media site. In the simplest case, keywords can be determined using word weighting methods. TF-IDF is one of the most popularly used and robust word-weighting methods, combining the effect of term frequency with inverse document frequency to determine the importance of words [9]. This method automatically gives a low value to common words like pronouns and prepositions that are normally neither relevant to a summary nor should be counted as key phrases. On the other hand a word with a high term frequency in the text we are trying to summarize and a low inverse document frequency would give a high TF-IDF value and thus identify a word that is important but not found in many different texts.

2.2 Pilot test

A pilot test with the baseline system is done to get a hands-on experience and benchmark with the performance of such systems, mainly the usefulness of the extracted tags.

What is considered a useful tag or a bad tag is mostly subjective and as such the only way of finding out whether the extracted tags are useful is to ask the users. The pilot test was set up so that each user gets twenty extracted tags and is asked to rate

the tags as useful, neutral or not useful. The system is limited to handling English language so only profiles with content mostly in English are considered valid.

The pilot test was conducted among international students and researchers and collected helpful feedback and observations concerning the type of useful tags we may use and system performance issues. In total we had 42 tests. Test results from non-English speakers were rejected from statistical analysis. This leaves the number of valid tests to 21. On average the percentage of useful tags was 43%, with highest at 55% (11 out of 20 tags) and lowest at 25% (5 out 20 tags).

In addition to interests and hobbies such as “cooking”, “biking” and “reading”, some other types of tags were also perceived as useful by the users. Those tags fall mostly into two categories: adjectives and names. Examples of positive adjectives would be “stylish” and “happy”. Names of people and locations were also perceived as positive. We can assume that the people are closely related or close friends if they are perceived as positive and that locations hold some special meaning for them to be perceived as positive tags. Such information could be linked to which city the person lives in, where the person grew up, or simply some place where the person recently has been.

While the system seems to work quite well on English profiles, it almost always failed to extract relevant tags from multilingual profiles. Other observations include the need for a social media specific stop word list, which could include slang and shortened words. Sentence ordering is not reliable in social media summarization. We expect sentence extraction to give a less than satisfactory result in the social media context. Keyword extraction takes into account neither sentence ordering nor length. As such, keyword extraction is perhaps more suitable to the analysis of social media content than a sentence based summary.

3 Extension of the baseline system

The pilot test showed us some obvious shortcoming in the system. As a further step, we want to be able to parse multilingual profiles so that finding test users becomes easier. We also want to look at the possibility of increase the accuracy of the extraction. So we try to extend our system further in three different areas. We try to increase the accuracy of the system by techniques that will help target relevant text portions in a profile. We implemented Named Entity Recognition to find names and locations of significance but not found through previous methods. We experiment with adding a predefined hobby and interest dictionary in case any hobbies or dictionaries receive low significance score in the extraction. Lastly we implement support for multiple languages through online translation. The extended approach is illustrated in Figure 2.

3.1 Targeting relevant text portion

Here we discuss ways to keep input data relevant, and how we can increase the significance of parts of the data to make it more relevant to the subject of the extraction.

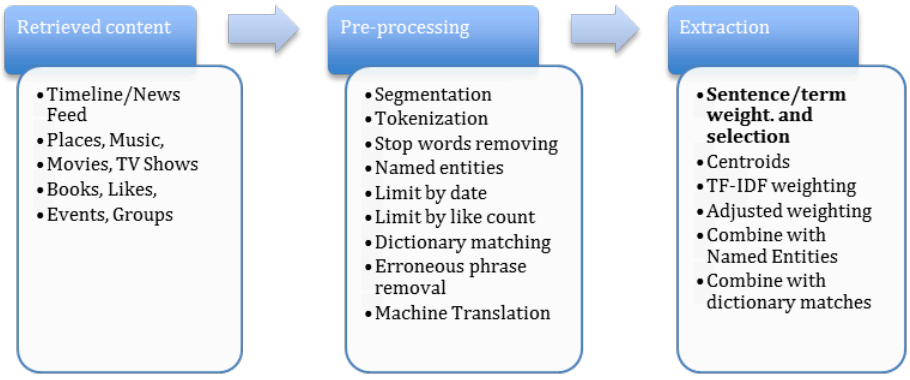

Figure 2 Determining the most representative terms for hobbies and interests in the extended system.

In addition to the TF-IDF weighting of word informative-ness, Luhn [10] suggested taking into consideration the positioning of sentences in texts. A sentence at the beginning of a paragraph or at the end of a text has a higher chance of being of high importance in the text [10]. However, this is not necessarily true for texts gathered from social media sites due to the following reasons: 1) the information will be structured according to whatever database model the company uses, 2) the information will be a collection of different areas such as personal information, communication with friends, interests and other issues which do not necessarily have anything in common, and 3) the data in question can include dialects, multiple languages, intentionally misspelled words etc.

As sentence ordering cannot be usefully taken into account in analyzing Facebook content, we suggest firstly: publishing dates, and secondly: site specific counters, as new criteria for determining relevant data. By limiting publishing to recent dates we can either leave out old information from social media profiles or decrease the significance of TF-IDF values for posts older than the specified date by a pre-determined factor. By using site-specific counters, for example increasing the significance of content that has received a higher amount of "likes" on Facebook, we can increase the content relevance. In addition, we can also try to make good use of semi-structured nature of pages such as those on Facebook and target content under different interest categories such as Places, Music, Movies, Books, Events, and Groups.

3.2 Dictionary based methods and machine translation

Posts containing several languages or non-English languages need to be translated for the extraction to work properly. As mentioned earlier social media content can consist of several different languages between posts but also within the same posts.

Dictionaries can be used in two ways to improve the results of our extraction. The first approach is to combine support for multiple languages through machine translation and dictionaries. Clark & Araki [7] introduces a system called Casual English Conversation System (CESC) to detect abbreviations, misspellings, punctuation omissions, non-dictionary slang, emoticons, wordplay, and censor avoidance. The approach is based on having databases/dictionaries and matching the text to the phrases in the database/dictionary and replacing the matched erroneous word with the grammatically correct version. [7]

The second way to use dictionaries is to directly extract hobbies and/or interests from social media content. This approach implies that we need dictionaries with the relevant hobbies and interests predefined. With the term weighting approach, new hobbies and interests will be hard to get identified.

The drawback of this dictionary approach is that, if a word is not in the interest/hobby dictionary, then it cannot be extracted. Unlisted hobbies and interests could never occur in an extraction based solely on this approach. The approach is easy to set up but hard to maintain as new hobbies and interests need to be added to the dictionaries to keep up to date.

To extend the capabilities of our system we use an approach that combines dictionaries for hobbies and interests with machine translation and extraction of terms with high significance. The result consists of all existing hobbies, interests, as well as words not registered in the dictionaries.

We use the freely available resource Yandex translation [11] to translate content into English. A separate stop word list is not needed for the extra languages since stop words will be translated to English and then removed by the English stop word list.

We continue by creating dictionaries for abbreviations, slang and intentionally misspelled words to supplement the system. An abbreviation is a shortened word, for instance the word “year” has an abbreviation “Yr.”. A typical example of an intentionally misspelled word would be writing “u” instead of “you”. Slang are words with the same meaning as another word but are not found in standard dictionaries, an example of an English slang word is “aggro” which often means the same as “angry” or “aggravate”. Such dictionaries can be either hand gathered or automatically gathered from the social media site if the structure of the database supports it. For instance, one approach would be to gather learning data from Facebook profiles. The information in question would be from the “like” and “group” tags.

When extracting the information in our system we first remove the abbreviation, slang and misspelled words by going through the dictionaries and match them to the profile that is being parsed. After that we translate the text into English. When we have the profile text in English we can use our interest and hobby dictionaries to supplement the TFIDF extraction. Lastly we decide how to order the extracted results. The options for ordering we have are; according to TFIDF significance, by publishing date, and/or by site-specific counters.

3.3 Named entity recognition

Named Entity Recognition (NER) is used to locate and classify names in texts [12]. While the name of the owner of a Facebook profile is not interesting when extracting hobbies and interests, there are still other named entities that could add value to the extraction. Recent work by Liu et al [13] includes semi-structured methods for finding Named Entities in Twitter "tweets". Tweets are not structured in any way and limited training data is available. This means that in order to reliably be able to extract Named Entities from tweets a semi-supervised method is created [13].

Hasegawa et al [12] presents an unsupervised approach to how we can link several Named Entities together. The approach works in the following way; first they Tag Named Entities in the text with a state-of-the-art NER tagger, then they pair Named Entities and look for similarities among the found pairs, lastly they cluster paired Named Entities and label them [12]. Liu et al [13] has also done research on Named Entity Recognition in Twitter "tweets".

When doing extraction of Named Entities in a Facebook profile we can take advantage of the semi-structured data. Each interaction on Facebook is linked to a person, and since we know which person we are extracting data from we can limit the Named Entities in an appropriate way. If we consider Named Entities as a possible part of the extracted words we would need to limit them so that only Named Entities that are relevant to the owner of the profile should be considered. The approach we end up with is a modified version of what Hasegawa et al [12] did for unstructured data, combined with Named Entities directly extracted from certain categories.

Since we are focusing on extracting data that is relevant to the profile owner we can directly extract Named Entities from the Facebook categories Groups and Pages. Groups consist of people that share a common interest; a user can share updates, photos and documents with other people in the group. Pages can be a place, company, institution, organization, Brand, Product, Artist, Band, Public Figure, Entertainment, Cause, or Community. Each profile is only linked to the Groups and Pages that the user him- or herself has decided to be linked to it. This means that Named Entities from these categories are linked to the user with a high probability. For the rest of the data we use the structure to pair Named Entities with the creator of the post, message or comment. Lastly we remove all unwanted Named Entities we have found in the profile from the list of Named Entities.

To be able to order the Named Entities according to relevance we can then increase the significance of the Named Entities so that they appear higher in the TF-IDF weighting. Another approach to sorting Named Entities by relevance is to find the highest weighted TF-IDF word that is linked to each Named Entity and sort them according to these TF-IDF values.

4 Conclusion and future work

In this study we investigate how to analyze people's social media profiles to extract hobby and interest information. We developed a baseline system that applies heuristic rules and TF-IDF term weighting method in determining the most representative terms indicating hobbies and interests. Our pilot test with limited amount of English-dominant user profiles shows 43% average useful tags, with highest at 55% and lowest at 25%.

The baseline system was extended to include new functionality to help set limits on the scope of relevant content, extract Named Entities, use of predefined dictionaries to identify even low-scoring hobbies and interests, and use of machine translation to handle content in multiple languages. When dealing with social media analysis it becomes important to support multiple languages as the extractions can fail or falsely portray unrecognized words as being of higher significance than they should be. When translating social media content we also need to take into account that people not necessarily follow grammatical rules.

Key word extraction seems more useful to the analysis of social media content than a sentence based summary, due to the fact that social media content is not structured as professional texts. Combining keyword extraction with predefined dictionaries and Named Entity Recognition gives us a broader scope. Named Entity Recognition becomes effective when combining state-of-the-art tools with the semi-structured architecture of for example a Facebook profile. Predefined dictionaries supplement the extraction by including lower-scoring words that still might have a high personal significance. An alternative to keyword extraction would be to extend the baseline system further to n-gram based weighting and topic models. In addition, LDA (Latent Dirichlet Allocation) topic modeling methods [14] could help us in identifying topics embedded in texts.

References

1. Agichtein, E., Castillo, C., Donato, D., Gionis, A., & Mishne, G. Finding high-quality content in social media. Proceedings of the international conference on Web search and web data mining (pp. 183-194). 2008. ACM.
2. Statisticbrain. 2014. Retrieved 16.01.2014 from http://www.statisticbrain.com/facebook-statistics/.
3. Delen D., Demirkan H. Data, information and analytics as services. Decision Support Systems, Volume 55(1), pp. 359-363. 2013.

4. Shamma, D. A., Kennedy, L., & Churchill, E. Summarizing media through short-messaging services. Proceedings of the ACM conference on computer supported cooperative work. 2010.
5. Zhao, W. X., Jiang, J., He, J., Song, Y., Achananuparp, P., Lim, E. P., & Li, X. 2011. *Topical keyphrase extraction from Twitter.* Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1 (pp. 379-388). Association for Computational Linguistics.
6. Yang, X., Ghoting, A., Ruan, Y., & Parthasarathy, S. A framework for summarizing and analyzing twitter feeds. Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining (pp. 370-378). 2012. ACM.
7. Clark, E., & Araki, K. Text normalization in social media: progress, problems and applications for a pre-processing system of casual English. Procedia-Social and Behavioral Sciences, 27, 2-11. 2011.
8. Bertoldi, N., Cettolo, M., & Federico, M. Statistical machine translation of texts with misspelled words. Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (pp. 412-419). 2010. Association for Computational Linguistics.
9. Salton, G., & Buckley, C. Term-weighting approaches in automatic text retrieval. Information processing & management, 24(5), 513-5. 1988.
10. Luhn Hans Peter. The automatic creation of literature abstracts. IBM Journal of research and development, 2(2), 159-165. 1958.
11. Yandex translation API. 2014. Retrieved from http://api.yandex.com/translate/
12. Hasegawa, T., Sekine, S., & Grishman, R. Discovering relations among named entities from large corpora. Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (p. 415). 2004, July. Association for Computational Linguistics.
13. Liu, X., Zhang, S., Wei, F., & Zhou, M, June. Recognizing Named Entities in Tweets. 2011. ACL (pp. 359-367).
14. Blei, David M., Ng, A. Y., & Jordan, M. I. Latent dirichlet allocation. Advances in neural information processing systems. pp. 2001. 601-608.