Clustering Facebook for Biased Context Extraction

Valentina Franzoni[1,2(✉)], Yuanxi Li[3], Paolo Mengoni[1,4], and Alfredo Milani[1,3]

[1] Department of Mathematics and Computer Science, University of Perugia, Perugia, Italy
franzoni@dis.uniromal.it, paolo.mengoni@unifi.it, milani@unipg.it
[2] Department of Computer Control and Management Engineering, University of Rome La Sapienza, Rome, Italy
[3] Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, Hong Kong
csyxli@comp.hkbu.edu.hk
[4] Department of Mathematics and Computer Science, Florence University, Florence, Italy

Abstract. Facebook comments and shared posts often convey human biases, which play a pivotal role in information spreading and content consumption, where short information can be quickly consumed, and later ruminated. Such bias is nevertheless at the basis of human-generated content, and being able to extract contexts that does not amplify but represent such a bias can be relevant to data mining and artificial intelligence, because it is what shapes the opinion of users through social media. Starting from the observation that a separation in topic clusters, i.e. sub-contexts, spontaneously occur if evaluated by human common sense, especially in particular domains e.g. politics, technology, this work introduces a process for automated context extraction by means of a class of path-based semantic similarity measures which, using third party knowledge e.g. WordNet, Wikipedia, can create a bag of words relating to relevant concepts present in Facebook comments to topic-related posts, thus reflecting the collective knowledge of a community of users. It is thus easy to create human-readable views e.g. word clouds, or structured information to be readable by machines for further learning or content explanation, e.g. augmenting information with time stamps of posts and comments. Experimental evidence, obtained by the domain of information security and technology over a sample of 9M3k page users, where previous comments serve as a use case for forthcoming users, shows that a simple clustering on frequency-based bag of words can identify the main context words contained in Facebook comments identifiable by human common sense. Group similarity measures are also of great interest for many application domains, since they can be used to evaluate similarity of objects in term of the similarity of the associated sets, can then be calculated on the extracted context words to reflect the collective notion of semantic similarity, providing additional insights on which to reason, e.g. in terms of cognitive factors and behavioral patterns.

Keywords: Word similarity · Semantic distance · Artificial intelligence · Data mining · Collective knowledge · Knowledge discovery

O. Gervasi et al. (Eds.): ICCSA 2017, Part I, LNCS 10404, pp. 717–729, 2017.
DOI: 10.1007/978-3-319-62392-4_52

1 Introduction

Facebook comments can be elicited by the aggregation of users in homophily [23] communities, e.g. by interest or opinion. We start from the observation that users can become polarized comment after comment, where they comment expressing similar concepts or with respect to a similar level of abstraction. Besides the preferential attachment approach [23] in fact, users often comment the main topic using the same use cases. For example, in the domain of information security, where a previous comment asks how to solve a problem, other users will probably seek help and create questions about the same problem, because they trust the source (who can be a previous commenter, or the user/page posting the main post) and they think to have the same problem. In information technology it also happens, as every computer scientist knows, in the well-known "fix-my-PC" problem. In Facebook, previous commenters can reinforce, and then drive, the polarization on particular sub-topics. Such sub-topics, containing in most cases an information bias, will often be off-topic with respect to the main post topics. In our work, we propose a process to separate clusters of *in-topicness*, where concepts underlying the content of comments are grouped by similarity with the concepts underlying the main topic. Experimental evidence, evaluated by human common sense, shows that such sub-topics form sub-contexts. In this work, posts and comments are extracted from the Facebook graph and are preprocessed with basic Natural Language Processing techniques [13]. The obtained bag of words is considered a set of candidate topics for sub-contexts. Semantic path-based WordNet distance [12] Leacock-Chodorow similarity [22] and Wu-Palmer similarity [20] are calculated, by means of the hierarchy of an ontological knowledge base, e.g. WordNet [1] where experiments have been implemented using path-based distances between pairs of term pairs (word1 from the main topic, word2 from each comment) for computation simplicity, but can be exploited also on Web-based semantic measures. The proposed approach can be applied to different distances in a social or collaborative taxonomy (e.g. Wikipedia [6, 7] Linked Data [24]). Preprocessed data, augmented with the similarity values, are then submitted to a clustering algorithm (e.g. Expectation-Maximization or simple K-means [25]) to obtain the sub-context clusters, that we validate by human common sense as a preliminary analysis. Since clusters are linkable to the same third party knowledge base (in our case, WordNet), in which the content similarity is calculated, a further evaluation can be done by referring to word-to-word semantic distance, or validating already accepted tagged data sets, where clusters can be compared to class tags to which a word pertains, or not.

The exploration of social networks or Web content using their semantic meaning is a consolidated modern approach to information extraction. The similarity measurement between documents and text has been extensively studied for information retrieval and Web-based measures [11, 12].

Content Based Image Retrieval (CBIR) [3, 18] enables satisfactory similarity measurements of low level features. However, the semantic similarity of deep relationships among objects is not explored by CBIR or other state-of-the-art techniques in Concept Based Image Retrieval and artificial intelligence. A promising idea is that it is possible to generalize the semantic similarity, under the assumption that semantically

similar terms behave similarly [4, 17, 19, 26]: the features of the main semantic proximity measures used in this work can be used in group similarity [27, 28] as a basis to extract semantic content, reflecting the collaborative change made on the web resources.

The example provided in Fig. 1 shows the different similarity recognition by humans and computers. Humans always have some bias [CIT SCIENCE FEB 2017] because of their cultural, educational or formation, besides the pure opinion that can be expressed in textual contributions to social or collaborative networks. Such a bias is a personal or community-based direction that will drive and shape opinions of other users participating to the same community, or potential ones. Such a bias is an important characteristic of the human being, and when politically wrong, it should be fixed by formation, not by filtering. With these premises, the most common problem of algorithms for automated tagging or context extraction is that they suffer to be domain-dependent. In particular machine learning approaches, which is the one with the best performance in many cases, suffers of the well-known problem of over-fitting. Now, it has been proved that semantics derived automatically from language corpora contain human-like biases: as quickly as it can learn, a machine learning process can amplify a bias. For instance, the pleasantness of a flower or unpleasantness of an insect can depend on cultural basis, but pushing too much the association between such accepted biases can lead over a racist threshold that, if generated by machines following human biases, is not acceptable by human politically-correct behavior.

In this point of view, is thus important that such biases are represented, being a content that will objectively shape opinions and cannot disappear in the analysis, but are not amplified, being considered a negative element. In other words, in this approach algorithms should not have opinions. The approach proposed in this paper is less domain-dependent, and does not pertain to that class of algorithms that can amplify the human bias, therefore can be preferred to machine learning, depending on the final goal, even when machine learning may have comparable or better results, which usually happens only in particular domains.

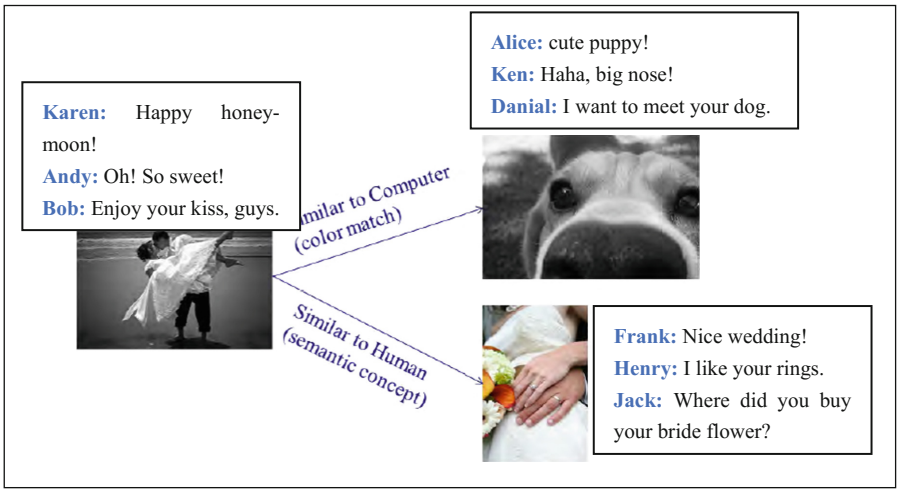

Fig. 1. Image similarity discovery comparison between computer and human

2 Related Work

2.1 WordNet Similarity

WordNet [1], is one of the applications of semantic lexicon propose for the English language and is a general knowledge base and common sense reasoning engine. Recent researches [2] on the topic in computational linguistics has emphasized the perspective of semantic relatedness of two lexemes in a lexical resource, or its opposite, semantic distance. The work in [12] brings together ontology and corpora, defining the similarity between two concepts *c1* and *c2* lexicalized in WordNet, named *WordNet Distance (WD)*, by the information content of the concepts that subsume them in the taxonomy. Then [27] proposes a similarity measure in WordNet between arbitrary objects where *lso* is the lowest super-ordinate (most specific common subsumer):

$$d(c_1, c_2) = \frac{2 \times \log p(lso(c_1, c_2))}{\log p(c_1) + \log p(c_2)} \tag{1}$$

The advantage of a WordNet similarity (where, results being normalized in a range [0, 1], similarity = 1 − distance) is to be based on a very mature and comprehensive lexical database, which provides measures of similarity and relatedness: WordNet, in fact, reflects universal knowledge because it is built by human experts; however, WordNet Distance is only for nouns and verbs in WordNet, but it is not dynamically updated. In Fig. 2, the "is a" relation example can be seen from [12].

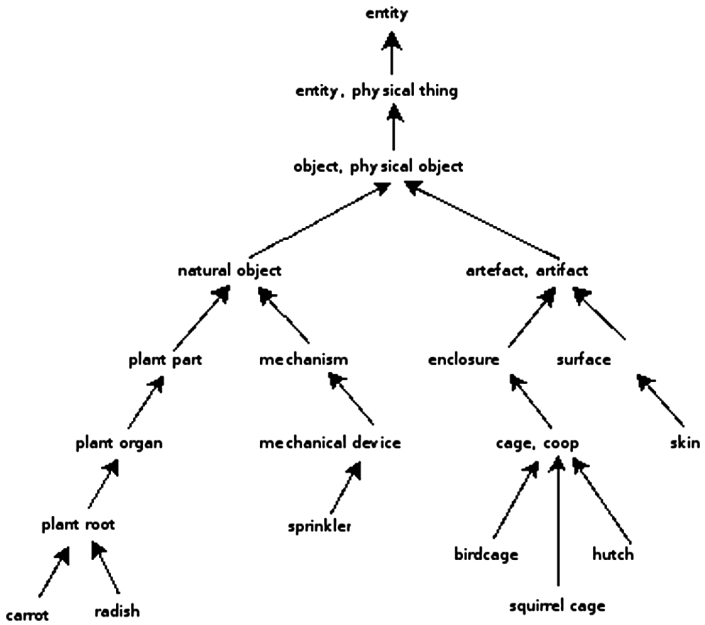

Fig. 2. WordNet "is a" relation example

2.2 Wikipedia Similarity

WikiRelate [5] was the first research to compute measures of semantic relatedness using Wikipedia. This approach takes familiar techniques that had previously been applied to WordNet and modified them to suit Wikipedia. The implementation of WikiRelate follows the hierarchical category structure of Wikipedia.

The *Wikipedia Link Vector Model (WLVM)* [6] uses Wikipedia to provide structured world knowledge about terms of interest. The probability of WLVM is defined by the total number of links to the target article over the total number of articles. Therefore, if t is the total number of articles within Wikipedia, the weighted value w for the link $a \rightarrow b$ is:

$$w(a \rightarrow b) = |a \rightarrow b| \times \log(\sum_{x=1}^{t} \frac{t}{|x \rightarrow b|}) \tag{2}$$

where a and b denote the search terms.

Among the approaches that use the hyperlink structure of Wikipedia rather than its category hierarchy or textual content, there is also the Heuristic Semantic Walk [26], that makes use of a search engine as a third-party knowledge base (e.g. Bing, Google) on which to calculate a Web-based similarity used as heuristic to drive a random walk. Wikipedia similarity reflects relationships as seen by the user community [7], which is dynamically changing as links and nodes are changed by the users collaborative effort. However, it only can apply to knowledge base organized as networks of concepts.

2.3 Flickr Similarity

Flickr distance (FD) [8] is another model for measuring the relationship between semantic concepts, in visual domains. For each concept, a collection of images is obtained from Flickr, based on which the improved latent topic-based visual language model is built to capture the visual characteristics of the concept. The Flickr distance between concepts *c1* and *c2* can be measured by the *square root of Jensen-Shannon divergence* [9, 15] between the corresponding visual language models, as follows:

$$D(C\_1,C\_2) = \sqrt{\frac{\sum_{i=1}^{K}\sum_{j=1}^{K} D_{JS}(P_{Zi}C_1 | P_{Zj}C_2)}{K^2}} \tag{3}$$

where

$$D_{JS} = (P_{Zi}C_1 | P_{Zj}C_2) = \frac{1}{2} D_{KL}(P_{Zi}C_1 | M) + \frac{1}{2} D_{KL}(P_{Zj}C_2 | M) \tag{4}$$

K is the total number of latent topics, which is determined experimentally. P_{Zi} C_1 and P_{Zj} C_2 are the trigram distributions under latent topic $z_i c_1$ and $z_j c_2$ respectively,

with M representing the mean of P_{Zi} C_1 and P_{Zj} C_2. The FD is based on Visual Language Models (VLM), which is a different concept relationship respect to WordNet Similarity and Wikipedia Similarity.

2.4 Context-Based Group Similarity

Set similarities in images [9, 10, 27], emotions [28] and, in general, web entities, can be calculated by means of underlying pair-based similarities with semantic proximity, based on user-provided concept clouds. A semantic concept cloud related to a Web object (e.g. image, video, post) includes all the semantic concepts associated to or extracted from the object. Typical sources for semantic concepts are tags, comments, descriptors, categories, or text surrounding an image. As shown in Fig. 3, Image I_i and Image I_j are a pair of images to be compared. T_{i1}, T_{i2},..., T_{im} are original user provided tags of image I_i, while T_{j1}; T_{j2},.., T_{jn} are original user provided tags of image IJ.

Given DI_{ij} as the distance (or equivalently, the similarity) of image I_i and image I_j, we define the ***Group Distance (GD)***:

$$DI_{ij} = AVG2\{AVG1[SEL(dT_{im\to jn})], AVG1[SEL(dT_{jn\to im})]\} \quad (5)$$

where *SEL* could be the maximum *MAX,* the average *AVG* or the minimum *MIN* of *d*, the similarity calculated by algorithm (*Confidence* or *NGD* [15] or *PMI* [14]), as in Eqs. (6–9).

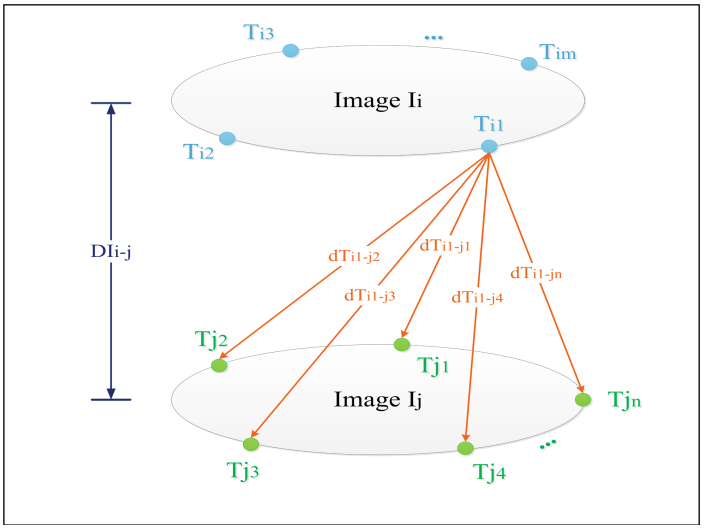

Fig. 3. Group similarity core algorithm

$$dT_{im} \rightarrow dT_{jn} = \begin{pmatrix} dT_{i1\rightarrow j1}, & dT_{i1\rightarrow j2}, & dT_{i1\rightarrow j3}, & \ldots & dT_{i1\rightarrow jn} \\ dT_{i2\rightarrow j1}, & dT_{i2\rightarrow j2}, & dT_{i2\rightarrow j3}, & \ldots & dT_{i2\rightarrow jn} \\ \ldots, & \ldots, & \ldots, & \ldots & \ldots \\ dT_{in\rightarrow j1}, & dT_{in\rightarrow j2}, & dT_{in\rightarrow j3}, & \ldots & dT_{in\rightarrow jn} \end{pmatrix}$$

$$dT_{im} \rightarrow dT_{jn} = \begin{pmatrix} dT_{j1\rightarrow i1}, & dT_{j1\rightarrow i2}, & dT_{j1\rightarrow i3}, & \ldots & dT_{j1\rightarrow im} \\ dT_{j2\rightarrow i1}, & dT_{j2\rightarrow i2}, & dT_{j2\rightarrow i3}, & \ldots & dT_{j2\rightarrow im} \\ \ldots, & \ldots, & \ldots, & \ldots & \ldots \\ dT_{jn\rightarrow i1}, & dT_{jn\rightarrow i2}, & dT_{jn\rightarrow i3}, & \ldots & dT_{jn\rightarrow im} \end{pmatrix} \quad (6)$$

$$AVG1\left[SEL(dT_{im\rightarrow jn})\right] = avg\left[SEL(dT_{i1\rightarrow jn}), SEL(dT_{i2\rightarrow jn}), , SEL(dT_{im\rightarrow jn})\right] \quad (7)$$

$$AVG1\left[SEL(dT_{jn\rightarrow im})\right] = avg\left[SEL(dT_{j1\rightarrow im}), SEL(dT_{j2\rightarrow im}), , SEL(dT_{jn\rightarrow im})\right] \quad (8)$$

$$AVG2 = AVGAVG1, AVG2 \quad (9)$$

3 Experiment

Information about Facebook post and comment similarity is extracted from raw data using a five-phases algorithm:

1. In the first phase Facebook post and comment data is harvested from public Facebook pages using an ad-hoc developed data pull app that have been registered on the social network.
2. Retrieved posts and comments are preprocessed to extract nouns.
3. Different ontology-based similarity measures are calculated on filtered nouns, where the distance between comments and the main topic are indagated.
4. Clustering is exploited on noun pairs augmented with similarity values.
5. Obtained clusters are visualized by a tag cloud and evaluated by means of human common sense.

Python, the Natural Language ToolKit [21], and TextBlob library are used to extract information, to analyze it using NLP techniques, and to compute word similarities.

3.1 Data Collection

Data are collected scraping the Facebook page *@Security*, which had (at time of experiments) over 9 million and 3 hundred thousand users. The access to Facebook data is allowed only to registered developers that write approved apps. The general policy on data access granted by Facebook include information from public Facebook pages or public posts written by normal users. To access private personal and post data

the user should install an app on his Facebook account and grant it specific permissions: in this case, apps can access all the data, for a limited time.

Based on this premise, our data extraction algorithm from Facebook uses public posts from the page, and comments related to each post. These data are requested to Facebook using the Facebook Graph API, a low-level HTTP interface to *node*, *edge* and *field* information, where nodes are Facebook objects (users, photos, pages, posts, comments, et cetera) connected through *edges* (photos in the page, and their comments) while *fields* are the specific information contained in nodes, i.e. attributes.

3.2 Preprocessing Phase

The extracted Facebook post and comment data have undergone a Part Of Speech (POS) tagging. Such preprocessing phase is needed to identify nouns, verbs, adjectives and other phrase components (Table 1).

Table 1. @Security page example of raw data

Topic: "Being safe online often starts with the developers who create the products we use every day. Today we're sharing tips for developers to write more secure code and help avoid security risks. #SID2017"
SampleComment1: "It would be nice to control if my post can be shared by other people. My sister in law shares all my posts"
SampleComment2: "I'd like to report a problem. I clicked on a photo story about Tomatoes and had an Attack on my computer. My security stopped the attack, and I quickly shut down. I went back to the site later and took a screen shot of the site, but never again clicked on it. I would like to show you the site, but don't see any place to post the photo. Please contact me."
SampleComment3: "How can I stop my friends seeing comments I make on other friend's posts who are not also their friends?"

3.3 Word-Level and Set-Level Similarity

After identifying the nouns contained in the post/comment, similarity between post and comment nouns is computed using two different strategies, each using three measures. The third-party knowledge base used for experiments is the lexical resource WordNet. In WordNet we identify the set of synonyms of nouns (i.e. synset) to which the word pertains, then we extract the first term included in the synset (as a synset name). Then, similarity (i.e. distance, by its inverse) is calculated by means of relations linking words, traversing the taxonomy through the hypernym hierarchy, i.e. "IS-A" relations.

The two implemented strategies differ on how Facebook comment features are extracted. The first technique uses a comment tag, i.e. one tag per comment, where the tag is a word used in the comment, using a set similarity. The other technique is based on exploiting the inner set similarities, calculating the pair distances between each of the nouns used in each comment and each word of the main post, i.e. the commented topic.

An adjacency matrix is then built on similarities, pair by pair, where similarities are the path-based WordNet distance [12], Leacock-Chodorow similarity [22] and Wu-Palmer similarity [20].

Using path similarity, a measure on how two words are similar is calculated based on the shortest path distance between the two terms found analyzing the hypernym relationship tree.

Leacock-Chodorow combines a taxonomy shortest path (i.e. length) between two associated word senses and the maximum taxonomy depth (D) using the following formula:

$$Sim_{Lch=} - log \frac{lenght}{2 * D} \tag{10}$$

Wu-Palmer similarity measure uses the taxonomy depth of two associated concepts (a and b) and the depth of the least common subsumer LCS (i.e. the nearest common parent concept (Tables 2 and 3)).

$$Sim_{w\&p=} \frac{2 * \text{depth(LCS)}}{\text{depth(a)} + \text{depth(b)}} \tag{11}$$

Table 2. @Security page example preprocessing: nouns extracted from text

Topic: "online, developers, products, day, Today, tips, developers, secure, code, security, risks, SID2017"
SampleComment1: "post, people, sister, law, shares, posts"
SampleComment2: "problem, photo, story, Tomatoes, Attack, computer, security, attack, site, screen, shot, site, site, place, photo, Please"
SampleComment3: "friends, comments, friend, posts, friends"

Table 3. @Security page example synset extraction from WordNet ontology

Topic: "developer, merchandise, day, today, tip, developer, code, security, hazard"
SampleComment1: "post, people, sister, law, share, post"
SampleComment2: "problem, photograph, narrative, tomato, attack, computer, security, attack, site, screen, shooting, site, site, topographic_point, photograph"
SampleComment3: "friend, remark, friend, post, friend"

3.4 Clustering Phase

Metrics of similarity provide data as distances in a Euclidean space. In general, any proximity measure can be used for clustering, even if it is not a metric, if the function following which the clustering algorithm will decide at each step where to include an evaluated point in a collection is defined (Table 4).

Table 4. @Security page example - similarity between each topic nouns' synset and each comment nouns' synset to be submitted for clusterization

Post noun	Post noun synset	Comment noun	Comment noun synset	Path sim.	LCH sim.	WUP sim.
Developers	Developer	Computer	Computer	0.091	1.240	0.444
Products	Merchandise	Computer	Computer	0.143	1.692	0.625
Day	Day	Computer	Computer	0.077	1.073	0.143
Today	Today	Computer	Computer	0.071	0.999	0.133
Tips	Tip	Computer	Computer	0.083	1.153	0.353
Developers	Developer	Computer	Computer	0.091	1.240	0.444
Code	Code	Computer	Computer	0.077	1.073	0.143
Security	Security	Computer	Computer	0.067	0.930	0.125
Risks	Hazard	Computer	Computer	0.091	1.240	0.286

The EM (Expectation-Maximization), as defined in [25] is an iterative algorithm for finding the maximum likelihood of estimated parameters, in statistical methods where the model depends on latent variables, e.g. from equations which cannot be resolved directly, or from data which were not observed, where the existence of such data can be assumed true. EM iteration rotates an expectation step (E), which iteratively calculates the expected likelihood on the current estimate of parameters, and a maximization step (M), which estimates which parameters maximize the expected likelihood, calculated in the E step until convergence, where updating the parameters does not increase anymore the likelihood.

K-means [25] is a clustering method to partition n observations into k clusters with the closest average (mean). The problem is computationally difficult (NP-hard), but algorithms exist, which make use of heuristics to converge quickly in a local optimum, similar to EM, through finishing steps.

3.5 Final Visualization

The human evaluation experiments have been held for the quality assessment of extracted sub-context, i.e. clusters. Experiments have been designed in a group of 12 experts, members of University of Perugia, from staff and students. Cloud tags related to sub-contexts clusters generated by the proposed algorithms from a pair of concept seeds extracted from Facebook comment pairs, have been submitted to the expert team. The experts have been asked to assess the relevance of the generated context on a range from 0 to 5 on a Linkert scale, by evaluating the context in the form of tags cloud where a term is shown in its cluster, with a size depending on its *in-topicness*. The clouds have been computed in three main of expertise for the different semantic proximity measures. In Fig. 4 we can see the tag cloud for the pair (Mars, Scientist) using a PMING-based HSW in (Wikipedia, Bing) from [11, 29] as an example. Tag clouds and dispersion graphs have been used, for visibility and readability issues. Tag clouds (see Fig. 4) basically

show a parameter (i.e. similarity value) at a time, where bigger terms correspond to the most similar words. These representations are well suitable to human evaluation. In our case, tag clouds represent each concept to be compared to the main topic.

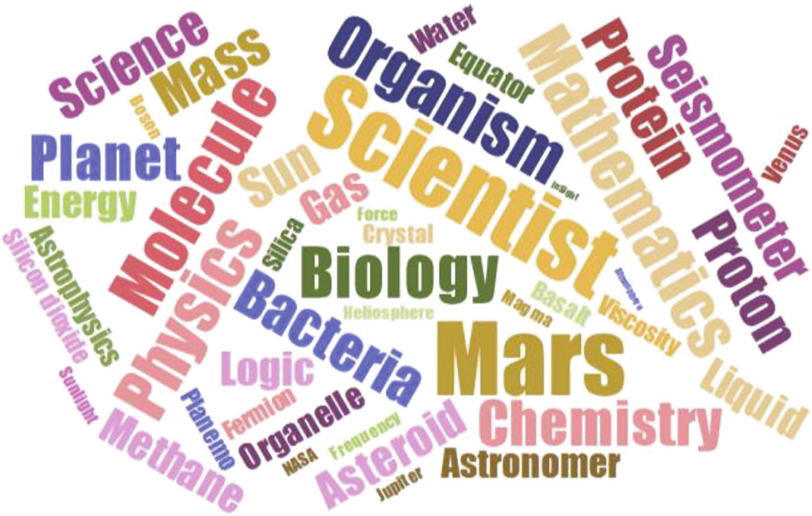

Fig. 4. Tag cloud for the pair (Mars, Scientist) using a PMING-based HSW in (Wikipedia, Bing)

4 Experimental Results and Discussion

In this work we introduced a method to investigate and identify the main context words obtained from Facebook posts and related user comments. The method is based on Natural Language Processing Part of Speech nouns extraction from sentences, similarity measurement using WordNet ontology, and clustering techniques.

Results show that clustering on frequency-based bag of words gives interesting results in the identification of topic contained in Facebook and it is more similar to human judgment than low level features comparison.

References

1. Miller, G.A.: Wordnet: a lexical database for english. Commun. ACM **38**(11), 39–41 (1995)
2. Budanitsky, A., Hirst, G.: Semantic distance in wordnet: an experimental, application-oriented evaluation of five measures. In: Proceedings of Workshop on WordNet and Other Lexical Resources, Pittsburgh, PA, USA, p. 641. North American Chapter of the Association for Computational Linguistics (2001)
3. Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence, pp. 448–453 (1995)

4. Lin, D.: An information-theoretic definition of similarity. In: Proceedings of the 15th International Conference on Machine Learning, pp. 296–304. Morgan Kaufmann (1998)
5. Strube, M., Ponzetto, S.P.: WikiRelate! computing semantic relatedness using Wikipedia. In: Proceedings of the Twenty-First National Conference on Artificial Intelligence. AAAI Press, July 2006
6. Milne, D., Witten, I.H.: An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In: WIKIAI 2008: Proceedings of First AAAI Workshop on Wikipedia and Artificial Intelligence, Chicago, IL, USA (2008)
7. Völkel, M., Krötzsch, M., Vrandecic, D., Haller, H., Studer, R.: Semantic Wikipedia. In: WWW 2006: Proceedings of the 15th International Conference on World Wide Web, New York, NY, USA, pp. 585–594. ACM (2006)
8. Wu, L., Hua, X.-S., Yu, N., Ma, W.-Y., Li, S.: Flickr distance. In: MM 2008: Proceedings of the 16th ACM International Conference on Multimedia, New York, NY, USA, pp. 31–40 (2008)
9. Enser, P.G.B., Sandom, C.J., Lewis, P.H.: Surveying the reality of semantic image retrieval. In: Bres, S., Laurini, R. (eds.) VISUAL 2005. LNCS, vol. 3736, pp. 177–188. Springer, Heidelberg (2006). doi:10.1007/11590064_16
10. Li, X., Chen, L., Zhang, L., Lin, F., Ma, W.: Image annotation by large-scale content-based image retrieval. In: Proceedings of the 14th Annual ACM International Conference on Multimedia, pp. 607–610 (2006)
11. Franzoni, V., Milani, A.: PMING distance: a collaborative semantic proximity measure. In: WI-IAT, 2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, vol. 2, pp. 442–449 (2012)
12. Leung, C.H.C., Li, Y., Milani, A., Franzoni, V.: Collective evolutionary concept distance based query expansion for effective web document retrieval. In: Murgante, B., Misra, S., Carlini, M., Torre, C.M., Nguyen, H.-Q., Taniar, D., Apduhan, B.O., Gervasi, O. (eds.) ICCSA 2013. LNCS, vol. 7974, pp. 657–672. Springer, Heidelberg (2013). doi:10.1007/978-3-642-39649-6_47
13. Manning, D., Schutze, H.: Foundations of Statistical Natural Language Processing. The MIT Press, London (2002)
14. Turney, P.D.: Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In: Raedt, L., Flach, P. (eds.) ECML 2001. LNCS, vol. 2167, pp. 491–502. Springer, Heidelberg (2001). doi:10.1007/3-540-44795-4_42
15. Cilibrasi, R.L., Vitanyi, P.M.: The Google similarity distance. IEEE Trans. Knowl. Data Eng. **19**(3), 370–383 (2007)
16. Li, Y.X.: Semantic image similarity based on deep knowledge for effective image retrieval. Research thesis (2014)
17. Franzoni, V., Mencacci, M., Mengoni, P., Milani, A.: Semantic heuristic search in collaborative networks: measures and contexts. In: WI-IAT (2), pp. 141–148 (2014)
18. Franzoni, V., Milani, A.: Heuristic semantic walk for concept chaining in collaborative networks. Int. J. Web Inf. Syst. **10**(1), 85–103 (2014). doi:10.1108/IJWIS-11-2013-0031
19. Franzoni, V., Mencacci, M., Mengoni, P., Milani, A.: Heuristics for semantic path search in Wikipedia. In: Murgante, B., et al. (eds.) ICCSA 2014. LNCS, vol. 8584, pp. 327–340. Springer, Cham (2014). doi:10.1007/978-3-319-09153-2_25
20. Wu, Z., Palmer, M.: Verbs semantics and lexical selection. In: Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics (1994)
21. Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. O'Reilly Media Inc., Sebastopol (2009)

22. Leacock, C., Chodorow, M.: Combining local context and WordNet similarity for word sense identification. WordNet: Electron. Lexical Database **49**(2), 265–283 (1998)
23. Bakshy, E., Rosenn, I., Marlow, C., Adamic, L.: The role of social networks in information diffusion. In: Proceedings of the 21st International Conference on World Wide Web, pp. 519–528. ACM (2012)
24. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: DBpedia: a nucleus for a web of open data. In: Aberer, K., et al. (eds.) SWC/ASWC 2007. LNCS, vol. 4825, pp. 722–735. Springer, Heidelberg (2007). doi:10.1007/978-3-540-76298-0_52
25. Franzoni, V., Milani, A.: A semantic comparison of clustering algorithms for the evaluation of web-based similarity measures. In: Gervasi, O., et al. (eds.) ICCSA 2016. LNCS, vol. 9790, pp. 438–452. Springer, Cham (2016). doi:10.1007/978-3-319-42092-9_34
26. Franzoni, V., Milani, A.: Heuristic semantic walk. In: Murgante, B., et al. (eds.) ICCSA 2013. LNCS, vol. 7974, pp. 643–656. Springer, Heidelberg (2013). doi:10.1007/978-3-642-39649-6_46
27. Franzoni, V., Leung, C.H.C., Li, Y., Mengoni, P., Milani, A.: Set similarity measures for images based on collective knowledge. In: Gervasi, O., et al. (eds.) ICCSA 2015. LNCS, vol. 9155, pp. 408–417. Springer, Cham (2015). doi:10.1007/978-3-319-21404-7_30
28. Biondi, G., Franzoni, V., Li, Y., Milani, A.: Web-based similarity for emotion recognition in web objects. In: UCC 2016, pp. 327–332 (2016)
29. Pallottelli, S., Franzoni, V., Milani, A.: Multi-path traces in semantic graphs for latent knowledge elicitation. In: ICNC 2015, pp. 281–288 (2015)