I.J. Modern Education and Computer Science, **2018, 6, 47-52**
Published Online June 2018 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijmecs.2018.06.06

Taxonomy Learning from Health Care Social Communities to Improve EHR Implementation

Zahia Marouf
King Abdulaziz University, Jeddah, Kingdom of Saudi Arabia
Email: zmarouf@kau.edu.sa

Sidi Mohamed Benslimane
LabRI-SBA Laboratory, École Supérieure en Informatique, Sidi Bel Abbes, Algeria
Email: s.benslimane@esi-sba.dz

Received: 12 February 2018; Accepted: 26 April 2018; Published: 08 June 2018

***Abstract*—**In this paper, we propose an approach to extract ontological structures from datasets generated by health care users of social networking sites. The objective of this approach is to exploit the user generated implicit semantics as a complement to more formalized knowledge representations. We aim for this latter to leverage the adoption level of the Electronic Health Record systems that are complaining from the shortage in standards and controlled vocabularies.

***Index Terms*—**Social networking sites, Ontologies, HER.

I. Introduction

In Web 2.0 Users are encouraged to add content, manage it, and share it with other users in an interactive and collaborative way. User content is created by users without the contribution of experts. This makes them relatively easy and rapid to build, conceptually simple, cheap, facile to use, and highly scalable.

Generating, or manipulating content is based on all kinds of terminologies varying from standard dictionary words and compound expressions to jargon and nonsense words. However, when a certain number of users agree on a term to describe or react to a content, this starts to look like a reasonable description of the content and forms a consensus about the term. This looks like the evolution of a natural language where a new word can be invented immediately and added to the vocabulary.

Electronic Health Record (EHR) is the documentation of a client's health status and healthcare in a secured digital format[16]. It is not simply an electronic version of the paper record as it offers different functionalities such as an integrated view of patient data, clinical decision support, clinician order entry, access to knowledge resources, and integrated communication and reporting support [17].

Even though the motivation of improved patient care and availability of medical data was present, health-care providers were hesitant to begin using this medical tool.

Specific reasons have been hypothesized for the lack of EHR implementation [19], one of them is the lack of standards in the healthcare domain as various programs are offering different features or using different standards and making the exchange of data not possible. Another reason is that EHR Always requires some process change by the provider and medical staff, as it brings a more rigid structure for entering information. Moreover Adapting to new standards for entering and manipulating information can be difficult initially and needs learning and training.

We believe that the insufficient involvement of users in the development of standards is a significant cause for the current shortage and the unsatisfying coverage found in domain ontologies [8][20][21].

In this paper, we adapt our approach to extract terminologies from folksonomies [6] to be used in social networking sites. The terminology is extracted from healthcare social networking site as a lightweight ontology. This later can be used as a communication vocabulary, a way to exchange and manipulate information between the different health information systems communicating with the EHR, and also to enrich the existing standards.

Our approach involves the healthcare providers in the terminology building, so it resolves the problems of the standards adoption and by the end, it improves the use and the quality of the EHR.

In section 2, we describe some related works. We outline the proposed approach and discuss the detailed steps in section 3. Sections 4 and 5 introduce an experimental methodology to evaluate the approach. Finally, section 5 concludes the paper and points directions for future work.

II. Related Works

The research area that extracts the collective intelligence of the collaborative tagging systems was vastly studied in recent years [6, 11, 12, 13, 14, 15]. However, few are approaches that try to extract ontological structures from social networking sites. We tried to gather here some of them that are more similar to our approach.

In [1], the authors propose an automated approach for Arabic slang lexicon extraction from microblogs for use in Arabic sentiment analysis.

Stefano et al [2] present a method for modeling twitter users supported by a hierarchical representation of their interests, which they call a Twixonomy.

Hoang et al [3] propose a model to represent the collection of microblogs into a knowledge base. In this approach, the authors suggest combining many tweets in the same context in order to resolve the tweets' sparsity problem (Given the size of a tweet, the information obtained by a single post is often very partial)

The authors in [4] investigate whether semantic relationships between entities can be learned from analyzing microblog posts published on Twitter and they found that Twitter is a suitable source as it allows for discovering trending topics with higher accuracy and with a lower delay in time than traditional news media.

In [5] the authors investigate a network-theoretic model called tweetonomy in order to study emerging semantics. They explore how the selection of the tweets (which they call Social Awareness Streams) can lead to different results. Their empirical findings demonstrate that different social awareness stream aggregations exhibit interesting differences, making them amenable to different applications.

Brambilla et al [9] propose an approach to discover low-frequency entities and establish their ontological properties. According to the authors, although low-frequency entities include emerging entities and may have a high impact in the future, they are not getting the same importance as most popular items in the process of ontological knowledge. Thus, they propose a method for discovering emerging entities by extracting them from social content.

Markus in [10] proposes an ontology enrichment pipeline that can automatically enrich a domain ontology using: data extracted by a crawler from social media applications, similarity measures, the DBpedia knowledge base, a disambiguation algorithm and several heuristics.

In [17], authors construct an ontology to model the semantics of social media streams, in particular, trending topics and public issues.

III. The Proposed Approach

In this section, we present our approach to learning hierarchical structures from healthcare social networking sites. It consists of the following three main steps: 1) representing microblogs as a graph of similarity between terms. 2) Clustering by grouping similar concepts. 3) Generating the hierarchy of concepts. The overall process is depicted in figure1.

A. Graph representation of microblogs

The approach collects data from selected hospital Twitter pages and represents them as a graph showing the association of concepts. This activity is performed in many steps:

a) The data are cleaned and filtered by removing stop words.
b) We recognize microblogs as a set of concepts that have their complex internal semantic relationships; furthermore, the sparsity of microblog messages (i.e., the limited length of messages) makes it challenging to extract semantics from them. For this reason, we use a concept mapping that matches the terms in each document to the entry terms in MeSH and then maps the selected Entry terms into MeSH Descriptors. If the word doesn't have any entry in Mesh, we map it to it most similar concept based on Levenshtein similarity. These descriptions are added to each post having the matched word.
c) After that, each post will be split into words and stored in a text file containing in each line one word, the ID of the user creating the post and the post ID.
d) Generality degree of each tag is generated based on FDU [6]. This measure needs to be adapted to fit the social network's different structure than the folksonomy. To this extent, we consider the post terms as tags, and we count the number of distinct users using this term in their posts as follow.

$$\forall\ t, \in T\ , u \in U, FDU(t) = \text{card}\{(u) \in U | t \in t_{up}\} \quad (1)$$

Where $t \in T$ are terms, $u \in U$ are users, $p \in P$ are posts t_{up} are terms used by the user "u" in the post p.

e) Building the association graph: This graph describes the associations between the different words used in microblogs. These associations are based on the similarity degree between these concepts. We use here the NCDU [6] similarity measure as it proved its quality against co-ocurrence and Cosine. NCDU also needs to be adapted to our new case as follow: $\forall$ t1, t2$\in$T , u$\in$U:

$$\text{NCDU}(t_1, t_2) = \frac{\sum_u w_u(t_1, t_2)}{\min(\text{FDU}(t_1), \text{FDU}(t_2))} \quad (2)$$

Where $\mathbf{w_u}(\mathbf{t_1}, \mathbf{t_2})$ are the values of the term-term per-user binary representations, described as follow:

$$\forall (t_1, t_2) \in T, u \in U, p \in P$$
$$w_u(t_1, t_2) = \begin{cases} 1 & \text{if } \exists\ t_{up} : (t_1, t_2) \in p \\ 0 & \text{else} \end{cases}$$

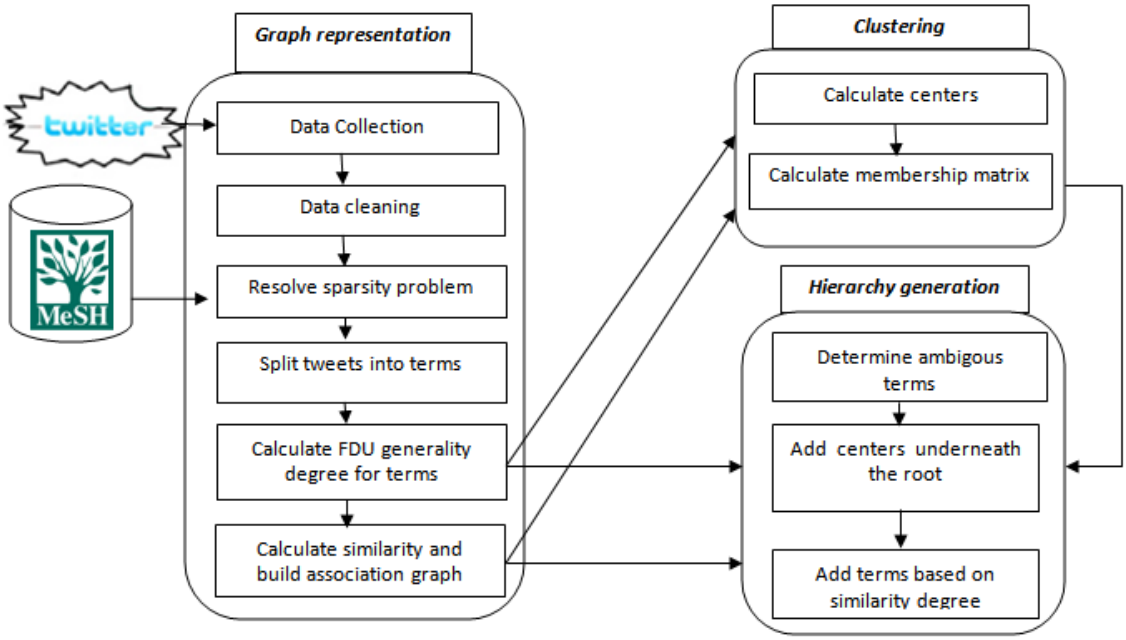

Fig.1. The proposed approach

B. *Clustering by combining similar terms*

In this step, we generate clusters by combining similar concepts. The number of clusters is not known a priory. Thus, we need an algorithm that starts with a set of terms and combines them based on their meanings. Furthermore, this algorithm is required to be able to assign ambiguous terms (same terms having different meanings) in more than one cluster depending on how many senses they. In [6] we proposed an algorithm with these characteristics and we evaluated its performance by comparing it against FCM. In this paper, we use the same algorithm with a small modification. The amelioration we made consists to prevent generating very small clusters by using a minimum size parameter (at least each cluster contains the center and one element). We attempt by this new version of the algorithm to ameliorate the quality of the generated clusters. Algorithm1 explains the clustering method.

C. *Hierarchy generation*

In this step, we induce a hierarchical organization scheme from the initially flat space that captures the semantics and the diversity of the shared knowledge.

In this paper, we adapt our algorithm for semantic identification to be used in any texts and not only folksonomies. Furthermore, we ameliorate the generated taxonomy by connecting ambiguous tags using "has_homonym". This way a search in the hierarchy with a given word will return all senses in case of ambiguity. The hierarchy generation is described in Algorithm2.

Algorithm 1 : Clustering

Inputs:

- NCDU[n][n] /*The similarity matrix of terms $t_1,\ldots,t_n$*/
- a list of terms $t_1,\ldots,t_n$ in descending order of their generality measured by FDU
- min_sim /* parameter for the threshold at which a word is considered belonging to a cluster */

Output:

- c /* the generated clusters' count*/
- U[n][c] /* the membership matrix of the terms*/

1: Add the first term in the generality list as the first center
2: Initialize the cluster count c=1
3: **For** all remaining tags *t* in the generality list
4: **If** NCDU[t][actualcenters] < minSim and there is at list a tag *t'* where NCDU[t][t']> min_sim **then** /* to ensure the cluster contains at list one element */
5: Increment the cluster count by one
6: Add t as the new center
7: **EndIf**
8: **EndFor**
9: For all centers and for all tags
10: Calculate the membership value u as NCDU[center][tag]
11: **Endfor**
12: **EndFor**
13: 16: return (c, U)

Algorithm 2: Hierarchy generation

Inputs:

- Lgenerality/*a list of terms t1,…,tn in descending vertex order of their generality measured by FDU */
-Membership matrix of terms U[c][n]
- min_sim/* parameter for the threshold at which a word is considered belonging to a cluster */
- c /* the centers count*/
- Centers[c] /* the centers set generated in the clustering step*/

Output:

- Hierarchy

1: **For** each term t /* First we determin ambiguous terms based on the clustering results*/
2: Nc[t] =card { c / u[t][c]> min_sim} /* here we calculate meanings's count as the number of clusters where t is included*/
3: Endfor
4: Hierarchy←< root>/* add root to the hierarchy*/
5: **for** i=1 to c **do**
6: is_a(root , centers[i]) /* add all centers to the hierarchy*/
7: Lgenerality. Remove(centers[i]);/* remove centers from the generality list*/
8: **end for**
9: **for** i=0 to $|L_{generality}|$ **do** / *for all terms in the generality list*/
10: Get the term number i 't_i ' from the generality list
11: **If** *t_i is an ambiguous tag ($Nc[t_i]>1$)* ***then***
12: *Is_a(t_i , root) /* add t_i underneath the root*/*
13: ***For*** *k=1 to $Nc[t_i]$ /* For all senses of t_i*/*
14: Identify the similar existing term t_j to t_i in the hierarchy
15: *t_i'= concat (t_i , (t_j)) /* here new concepts are generated by concatenating each ambiguous term with its different meanings.*/*
16: *is_a (t_i' , t_j) /* t_i' is added underneath t_j */*
17: has_homonym (t_i , t_i') /* t_i , t_i' are connected */
18: ***EndFor***
19: ***Else /**** t_i is not ambiguous*/
20: Identify the most similar existing term t_j to t_i in the hierarchy
21: *is_a (t_i , t_j) /* t_i is added underneath t_j */*
22: If no similar term exists for t_i **then**
23: *Is_a(t_i , root) /* add t_i underneath the root*/*
24: **End if**
25: **End if**
26: End for
27: Return (Hierarchy)

IV. Experiments

We perform a set of experiments in order to test the results of applying our approach to user-generated content other than folksonomies. The dataset is a set of tweets we got from the open source Healthcare Twitter Analysis Project[1]. In the following, we explain each experimentation step.

A. Graph representation of microblogs

In this step, we generate a list of terms ordered by their generality degree FDU, and a similarity matrix NCDU.

Tables II and III depict excerpts of these representations.

B. Clustering

As explained above, each term is assigned to one or more clusters depending on how many senses it can have.

Table I shows some examples of ambiguous terms detected by the algorithm.

Table I. An expert of discovered ambiguous terms with their context

Terms	Cluster centers
Fluid	liquid/behavior
Glucose	substance/measurement
Radiation	power/therapy
Support	device/ care

C. Hierarchy generation

In this step, we have generated a hierarchy of terms using the algorithm described in section III.C. Figure 2 illustrates an excerpt of this hierarchy.

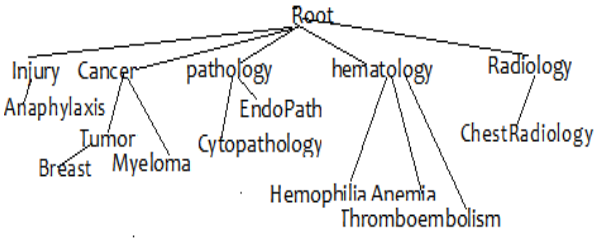

Fig.2. An excerpt of the generated hierarchy

V. Results

As for the overall approach, it was evaluated in depth in previous works and it proves its quality. However, the hierarchy generated in previous research was for English terms and for its evaluation, we compared it against manually built categorization schemes from WordNet and Wikipedia.

In this research, we evaluate excerpts from our generated hierarchy against healthcare ontologies generated from social media by a group of experts in the symplur project[2] (a healthcare social media analytics company). Although their ontologies are not mature enough, they can give us insight about the quality of our results using a measure such as F-measure. Figure 3 depicts the comparison results where a high score of F-measure is recorded for each ontology.

[1]http://healthcaredataanalysis.org/main/
[2] https://www.symplur.com/

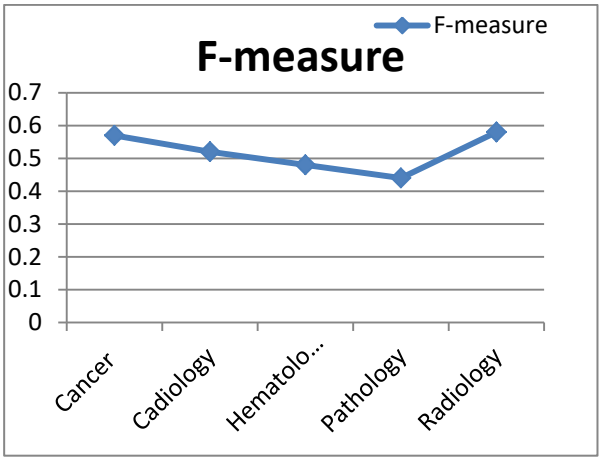

Fig.3. Evaluation

Despite these results, we need to compare it with more accurate healthcare ontologies. The work is in progress to find a good ontology for this purpose. Even though we can say that the preliminary results are satisfactory.

VI. Conclusion

In this paper, we have proposed an integrated approach to extract ontological structures from healthcare social networking sites. We proposed an adaptation to our previous approach based on the three-dimensional space of folksonomies to a simple text processing approach. This adaptation also considers the social dimension and it impacts on improving the quality of the resulting semantics.

For future work, rigorous evaluations need to be performed on the approach in order to get a more clear view of the approach performance in real applications.

Furthermore, a lack of specificity in the existing informatics taxonomies for terminology used to describe the growing field of health information technology (health IT) created the need for the development of a specialized taxonomy [7]. For this purpose a new ontology for health IT is planned as future work.

Table II. An excerpt of generality list

Term	Health	Heart	Cardio	Disease	Cancer	Cardiomyopathy
Generality degree	6561	5114	2030	890	720	265

Table III. An excerpt of the NCDU Matrix

	Leukaemia	BloodCancer	Disease	Cholesterol	Hepatitis
Leukaemia	0	0,86	0.52	0.21	0.04
BloodCancer	0,86	0	0.61	0.11	0,17
Disease	0,52	0.61	0	0,47	0,75
Cholesterol	1.21	0.11	0.47	0	0,007
Hepatitis	0.04	0,17	0.75	0,007	0

References

[1] Samhaa R. El-Beltagy, A Fully Automated Approach for Arabic Slang Lexicon Extraction from Microblogs, Proceedings of the 15th International Conference on Computational Linguistics and Intelligent Text Processing, April 06-12, 2014, Kathmandu, Nepal.

[2] Stefano Faralli, Giovanni Stilo, Paola Velardi: Automatic acquisition of a taxonomy of microblogs users' interests. J. Web Sem. 45: 23-40(2017).

[3] Hoang, ThiBich Ngoc and Mothe, Josiane. Building a Knowledge Base using Microblogs: the Case of Cultural MicroBlog Contextualization Collection. (2016) In: Conference and Labs of the Evaluation forum (CLEF 2016), 5 September 2016 - 8 September 2016 (Evora, Portugal).

[4] IlknurCelik, Fabian Abel, and Geert-Jan Houbenn. Learning semantic relationships between entities in twitter. In ICWE, 2011.

[5] Claudia Wagnerand MarkusStrohmaier. The Wisdom in Tweetonomies: Acquiring Latent Conceptual Structures from Social Awareness Streams, Semantic Search 2010 Workshop (SemSearch2010), in conjunction with The 19th International World Wide Web Conference (WWW2010), Raleigh, NC, USA, April 26-30, ACM, 2010.

[6] MaroufZahia, Sidi Mohamed Benslimane : Towards Ontological Structures Extraction from Folksonomies: An Efficient Fuzzy Clustering Approach. *IJIIT* 10(4): 40-50 (2014).

[7] Dixon, B. E., Zafar, A., & McGowan, J. J. (2007). Development of a taxonomy for health information technology. In Medinfo 2007: Proceedings of the 12th World Congress on Health (Medical) Informatics; Building Sustainable Health Systems (p. 616). IOS Press.

[8] Lamsfus, C., Alzua-Sorzabal, A., Martin, D., Salvador, Z., &Usandizaga, A. (2009). Human-centric ontology-based context modelling in tourism. In Proceedings of the international conference on knowledge engineering and ontology development, Funchal, Madeira, Portugal, October 6–8, 2009 (pp. 424–434).

[9] Brambilla, M., Ceri, S., Valle, E.D., Volonterio, R. Salazar, F.X.A.: Extracting emerging knowledge from social media. In: 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017. pp. 795–804 (2017).

[10] P. Monachesi and T. Markus. Using social media for ontology enrichment. In Extended Semantic Web Conference, pages 166–180. Springer, 2010.

[11] Mika, P.(2007).Ontologies are us: A unified model of social networks and semantics.In Journal of Web Semantics. 5(1), 5-15.

[12] Schmitz, P. (2006). Inducing ontology from Flickr tags. Collaborative Web Tagging Workshop, 15th WWW Conference, Edinburgh.

[13] Specia, L.& Motta, E. (2007). Integrating Folksonomies with the Semantic Web. In: 4th European Semantic Web Conference, ESWC,pp. 624-639, Innsbruck, Austria.

[14] Lin, H., Davis, J. &Zhou, Y. (2009).An integrated approach to extracting ontological structures from folksonomies. In Proceedings of the 6th European

Semantic Web Conference on The Semantic Web: Research and Applications, page 668. Springer.
[15] Heymann. P. & Garcia-Molina. H. (2006). Collaborative creation of communal hierarchical taxonomies in social tagging systems. Technical Report 2006-10, Computer Science Department.
[16] F. Ozair, N. Jamshed, A. Sharma, and P. Aggarwal, “Ethical issues in electronic health records: A general overview,” *Perspectives in Clinical Research*, vol. 6, no. 2, pp. 73–76, 2015.
[17] Tang PC, McDonald CJ. Electronic health record systems. In: Shortliffe EH, Cimino JJ, eds. Biomedical informatics: computer applications in health care & biomedicine. New York: Springer-Verlag; 2006:447–75.
[18] ZHOU Y, Li W, YUAN X, Zhang P. Ontology modeling of semantics in social media: Public issue knowledge base (PIKB) of the Weibo. Chinese Journal of Library and Information Science. 2014;1:002.
[19] Hamilton, B. (2011). Electronic health records. New York, NY: McGraw-Hill.
[20] Marco Alfonse, Mostafa M. Aref, Abdel-Badeeh M. Salem,"An Ontology-Based System for Cancer Diseases Knowledge Management", *IJIEEB*, vol.6, no.6, pp.55-63, 2014. DOI: 10.5815/ijieeb.2014.06.07.
[21] Iroju Olaronke, Ojerinde Oluwaseun, “An Ontology Based Remote Patient Monitoring. Framework for Nigerian Healthcare System", *International Journal of Modern Education and Computer. Science (IJMECS)*, Vol.8, No.10, pp.17-24, 2016.DOI:10.5815/ijmecs.2016.10.03.

Authors’ Profiles

Zahia Marouf received the Magister and Ph.D. degrees in computer Sidi Bel Abbes University- Algeria in 2010 and 2015 respectively. In 2011 she joined the Department of Economics and Management, University of Mascara, Algeria as an assistant professor. Since December 2015, she has been with the information system Department of King Abdulaziz University. She is a member of the Evolutionary Engineering and Distributed Information Systems Laboratory, EEDIS. In research, her current interests include ontology learning, information and knowledge management, Semantic Web.

Sidi Mohamed Benslimane is a full Professor at the Higher School of Computer Science, Sidi Bel-Abbès, Algeria. He received his PhD degree in computer science from Sidi Bel Abbes University in 2007. He also received a M.S. and a technical engineer degree in computer science in 2001 and 1994 respectively from the Computer Science Department of Sidi Bel Abbes University, Algeria. He is currently Head of Higher School of Computer Science, Sidi Bel-Abbès, Algeria. From 2001 to 2015, he was a member of the Evolutionary Engineering and Distributed Information Systems Laboratory, EEDIS. Actually, he heads the Research Team ‘Service Oriented Computing’; at LabRI-SBA Laboratory. His research interests include, semantic web, service oriented computing, ontology engineering, information and knowledge management, distributed and heterogeneous information systems and context-aware computing.

How to cite this paper: Zahia Marouf, Sidi Mohamed Benslimane, " Taxonomy Learning from Health Care Social Communities to Improve EHR Implementation", International Journal of Modern Education and Computer Science(IJMECS), Vol.10, No.6, pp. 47-52, 2018.DOI: 10.5815/ijmecs.2018.06.06

