A Health Information Recommender System: enriching YouTube Health Videos with Medline Plus Information by the use of SnomedCT terms

A. Rivero-Rodríguez[1], S. Th. Konstantinidis[1], C.L. Sanchez-Bocanegra[2], L. Fernández-Luque[1]
[1]*Norut, Tromsø, Norway*
[2]*Robot. & Comput. Technol. Lab., Univ. of Seville, Seville, Spain*
alejandro.rivero@norut.no, Stathis.Konstantinidis@norut.no, carlosl.sanchez@atc.us.es, luis.luque@norut.no

Abstract

Web 2.0 is the web of collaborating and sharing, where all users have the chance to create, publish and share content. Thus there are two important effects. There is an overload of the information on the web and the trustworthiness of the sources is uncontrolled. In the health domain, access to harmful information could be controlled. To reach this goal we propose a Health Information Recommender System to connect videos with trustworthy information from very trustful medical sources, such as Medline Plus. According to video's data, this system detects the main topic of the video and enriches it with information from very well-known resources. Evaluation results reveal that the method using SNOMED CT terms to identify relative information is the most appropriate as the main method of the Health Information Recommender System.

1. Introduction

Internet evolution makes new ways of understanding it. Web 2.0 has had a great impact on the quantity of information that is available on the Internet. Since everyone can post information, content quality cannot be ensured. That could be considered as a disadvantage for the health information.

During the last years there is a significant increase of online searches for health information[1], while the quality of online information is very heterogeneous. This could be considered as a public health concern [2], as far as it increases the complexity for most health consumers to discern when content is reliable or not.

Multiples initiatives have been promoted in order to facilitate health consumers to find reliable online information. There are a lot of example in which the focus has been toward the creation of quality labels and certifications for trusted websites [3-7].

To this extent, there were proposed many methods of controlling the access to unreliable information. A common one is to restrict social networks. A common one is to. A common one is to restrict social networks avoiding the publication of harmful or inappropriate information, as YouTube already does with sexual content. Fewer efforts proposed the use of semantics in order to provide better recommendation within social networks [8],[20], while other researchers suggest that users should avoid the use of search engines where misleading information can be ranked on the top, such as Bing or Google [9]. In social networks a set of relevant recommendation links can be provided. Therefore, we avoid that health consumers search information in non-trustworthy sources.

Another approach to recommend relative health information is to create predefined libraries of educational resources [10]. This approach has disadvantages, such as impossibility of dealing with the increasing quantity of online content. But, it ensures that trustworthy content is offered. Medline Plus is one of these health repositories for patients. It is a service offered by NLM of USA (National Library of Medicine) aiming to provide quality health information about diseases, treatments, etc.

In this paper, a method to enhance health videos from YouTube with content from Medline Plus is proposed and evaluated. The key factor is the inclusion of filters to identify medical-related terms in video's

titles, in order to make recommendations based on identified medical concepts by the use of SNOMED CT.

2. Background

Medline Plus [11] is a service offered by the National Library of Medicine of USA that aims to assist citizens in finding authoritative health information. This information is considered trustworthy and includes many different sections such as information about diseases, dictionaries with spelling and definitions of medical terms or publications. Medline Plus offers a web service to access the indexed information.

To this extent, SNOMED CT (Systematized Nomenclature of Medicine Clinical Terms) is a systematically organized computer collection of medical terms providing codes, terms, synonyms, etc. [12]. SNOMED CT provides a general terminology for electronic health records (EHRs) and contains more than 311,000 active concepts. SNOMED CT consists of concepts, descriptions and relationships. Each concept is associated with one or more descriptions, which are human readable terms, and information about the terms [21]. In this paper we consider as a SNOMED CT term any term that is human readable and exists in the description of a SNOMED CT concept.

These two concepts can provide the chance of enhancing health information in a very simple way. Researchers have tried to use Medline Services in order to link information from PHR and Medline Plus.

There have been many works and initiatives toward that topic. We find theoretical ways to adapt recommender systems to requirements of PHR systems, based on Semantic Network and Graph Theory [13]. Microsoft has developed a Personal Health Explorer, based on Semantics, to recommend content within Microsoft's Health Value [14]. Another example, the University of Murcia [15] has developed an application to offer trustworthy interesting information from Medline Plus to users according to an application that manages pills and treatment.

Why not to use big health information from Medline Plus to link from other kind of information? Recommender Systems for social network is a relevant field and many different algorithms have been published. These algorithms, such as those used by YouTube or other popular online networks, can be applied in order to achieve good recommendations [16] in the health domain. However, specific recommendation methods for health information can be better due to the development of Structured Health Information (e.g. SnomedCT).

In our experiment we try to explore recommendations for YouTube videos, based on the titles, by filtering the medical terms and offer links to Medline Plus, providing general information about the main matter of the video.

3. Method

In this paper recommendation system methods for health information are proposed and compared. During this study, only the title of the videos was taken into account. Figure 1 shows the general flow diagram of the proposed health information recommender system.

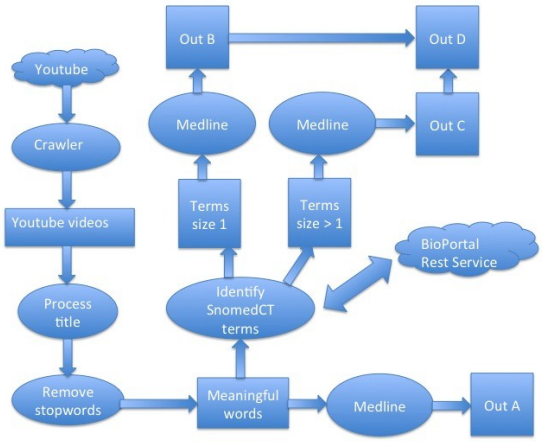

Figure 1: Flow diagram of the Health Information Recommender System

Crawler extracts 1000 videos from YouTube, from the twenty most popular American hospital channels, according to Bennett [17]. Once the video information is crawled, the system extracts exclusively the title of the videos. Furthermore, the aim of this experiment is to prove that filtering by medical terms can be positive in order to enhance with medical content.

Process of the title divides sentence into words, eliminating dots, spaces and other characters. The results are stored in an array, with one word in each cell. Furthermore stopwords are removed from this array. Stopwords are very general words that do not provide much meaning to the content, such as

prepositions, articles, or common verbs. As a result we have all the meaningful words ("Meaningful Words" in Figure 1).

"Medline" ellipsis in Figure 1 is a function that used the Medline Plus API, which search by the use of the "Meaningful Words" in the Medline Plus indexed information. This API provides back to the Health Information Recommender System a list comprised of the Medline Plus Information webpages.

"Identify SNOMED CT terms" ellipsis in Figure 1 is a function that used the Bioportal API [22], which identify whether a term is a SNOMED CT terms or not. The use of the parameter "isexactmatch" (set to off) was made in order to allow partiar matches to the queries.

Four different methods were implemented and compared. That provided for output sets of relative information:

Method A: This method used the Medline Plus API to provide Medline Plus Information webpages without taking into account whether the terms were medical related or not (Output A in Figure 1). This method could be considered as a generic recommendation system method.

Method B: This method was filtering the terms to identify whether or not they were SNOMED CT terms by the use of Bioportal API. This method was matching only one term of the title and not combination of more than one term as SNOMED CT terms. Then the method used Medline Plus API to provide Medline Plus Information webpages (Output B in Figure 1).

Method C: This method was filtering the terms to identify whether or not they were SNOMED CT terms by the use of Bioportal API. This method was matching terms of the title, which consisted of two or more words as SNOMED CT terms, being complementary to method B. hen the method used Medline Plus API to provide Medline Plus Information webpages (Output C in Figure 1).

Method D: Method B & C were compound and created method D in which all length of terms were tried to matched with a SNOMED CT term. The matched terms were used by the Medline Plus API to provide Medline Plus Information webpages (Output D in Figure 1).

Once the data process had been carried out, an evaluation method has been applied. This evaluation method includes two parts. The first one is a statistical process of data in order to measure the amount of related Medline Plus Information webpages per method. The second one is the quality evaluation. Five videos were extracted randomly out of the bigset of 1000 videos. Each video accompanied two lists of recommended Medline Plus Information webpages, corresponding to Methods A & D. Two expert users evaluated video recommendation deciding for each link whether the content of the recommendation and video were related.

4. Results

As explained above, we applied the methods A and D on a set of 1000 videos, extracted from the twenty most popular American hospital channels in YouTube.

Method A mean value of recommended Medline Plus Information webpages per video is much higher than the method D, as depicted in Figure 2:

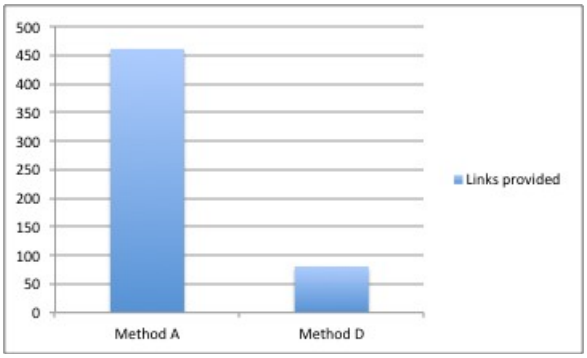

Figure 2: Number of recommended links by methods

For the qualitative evaluation of the users method A and D were selected.There were five random videos extracted for quality evaluation. The methods calculate the Medline Plus recommendation links. Those links were evaluated of their relativeness with the video content itself by two experts.

Hits rates indicate the percentage of relevant recommendation links. We calculated the hit rate of the method as the average of the hit rage of the method for the five selected videos, evaluated by the two users.

The hit rate for method A was 4.94% while for method D was 46.44%.

Table : Hit rates of Methods A & D

Hit rate A (%)	4.94
Hit rate D (%)	46.44

In the next figure both number of links and hit rate, are shown.

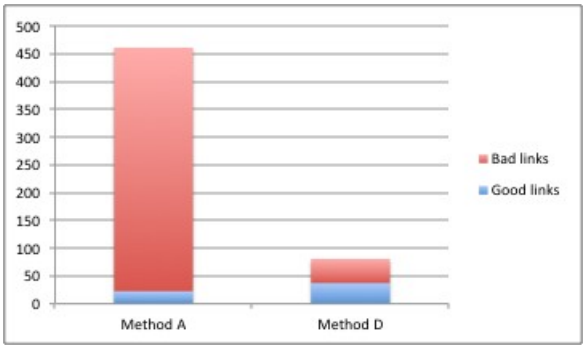

Figure 3: Recommendation links by methods. Unrelated and related links.

It is relevant to underline that Method D offers two advantages: on one hand the number of related links is higher and there are more relevant information for the users. On the other hand there are less unrelated links and consequently, it is much more probable to find information related to the video.

It is also very important to refer to the similarity of expert users decision on how similar were their evaluations. The rate of agreement to classify links was 96.9%, which mean decision criteria has been well defined.

5. Discussion & Future Work

Obviously Internet users already have Internet habits and changing them is extremely hard. It takes long time and lot of money to make people conscious of the matter of health quality in the web, when we consider all the webpages in Internet. For instance, if they use Google to search information about diabetes, it is very difficult to change their attitude significantly. In contrast the proposed Health Information Recommender System using method D could help users find related trustworthy formation within social networks.

The proposed method D as a Health Information Recommender System can be applied in Health Social Networks. Our aim is to enrich the information from health videos with trustworthy health information from other sources.

The results reveal that retrieving relative health information by the use of SNOMED CT terms is more accurate comparing with retrieving health information by querying including all the keywords. Results are far better using the method D (hit rate 46.44%) than applying the method A (hit rate 4.94%). This meets our expectations, since it was logical to think that extracting medical terms is enough to determine the main matter of videos.

In the limitations of our study could be considered that the recommendations are based exclusively in the titles, which sometimes are not very representative of the content (e.g. video's title is Linda's Story). Occasionally videos are just used as advertisement for hospitals.

Our future work will engage more metadata of the health videos to enrich the information about the video and therefore provide better Medline Plus links. Technologies such as the automatically abstraction of subtitles from YouTube health video content will further investigated in order to increase the quality of the detection of relative health information, providing better recommendations.

A closer look at the resulted data revealed that there were lot of unrelated links repeated in many videos. The reason was the appearance of some common medical terms (e.g. cancer) which exist in many related information. Our future plan is to include more advanced health stopwords filter, such as in [19].

Last but not least, even if the Medline Plus library contains high quality health information for patients, there is not enough information to cover all the topics introduced in the YouTube health videos. So different silos of related information should be identified.

Acknowledgement

This work has received support from pEducator project of the Tromsø Telemedicine Laboratory co-funded by the Norwegian Research Council, project 174934. This work received further support from the HelseTV project funded by the University of Tromsø.

6. References

[1] Kummervold PE, Chronaki CE, Lausen B, Prokosch HU, Rasmussen J, Santana S, et al. eHealth trends in Europe 2005-2007: a population-based survey. J Med Internet Res 2008;10(4):e42 [2] McLeod SD. The quality of medical information on the Internet. A new public health concern. Arch Ophthalmol. 1998 Dec;116(12):1663-5.
[3] HON's MARVIN project. HON Official website. [Access]
[4] MedCertain. Health Informatic Europe Official website. [Access]
[5] MedCircle Official Website [Access]
[7] Mayer MA, Karkaletsis V, Stamatakis K, Leis A, Villarroel D, Thomeczek C, Labský M, López-Ostenero F, Honkela T. MedIEQ-Quality labelling of medical web

content using multilingual information extraction. Stud Health Technol Inform. 2006;121:183-90.
[8] Breslin, J.; Decker, S.; "The Future of Social Networks on the Internet: The Need for Semantics," Internet Computing, IEEE , vol.11, no.6, pp.86-90, Nov.-Dec. 2007 doi:10.1109/MIC.2007.138
[9] Wang L, Wang J, Wang M, Li Y, Liang Y, Xu D Using Internet Search Engines to Obtain Medical Information: A Comparative Study. J Med Internet Res 2012;14(3):e74
[10] Fernandez-Luque L, Karlsen R, Vognild LK. Challenges and opportunities of using recommender systems for personalized health education. Stud Health Technol Inform. 2009;150:903-7.
[11] Naomi Miller, Eve-Marie Lacroix, Joyce E. B. Backus. MEDLINEplus: building and maintaining the National Library of Medicine's consumer health Web service. Link
[12] Ruch, Patrick; Gobeill, Julien; Lovis, Christian; Geissbühler, Antoine (2008). "Automatic medical encoding with SNOMED categories". BMC Medical Informatics and Decision Making 8: S6. doi:10.1186/1472-6947-8-S1-S6. PMC 2582793.
[13] Martin Wiesner and Daniel Pfeifer. 2010. Adapting recommender systems to the requirements of personal health record systems. In *Proceedings of the 1st ACM International Health Informatics Symposium* (IHI '10), Tiffany Veinot (Ed.). ACM, New York, NY, USA, 410-414. DOI=10.1145/1882992.1883053
[14] Morrell, T.G.; Kerschberg, L.; , "Personal Health Explorer: A Semantic Health Recommendation System," Data Engineering Workshops (ICDEW), 2012 IEEE 28th International Conference on , vol., no., pp.55-59, 1-5 April 2012 doi: 10.1109/ICDEW.2012.64
[15] UMU Norut PHR Powered by Indivo (video). http://www.youtube.com/watch?v=sz6CMAOY8jg
[16] James Davidson, Benjamin Liebald, Junning Liu, Palash Nandy, Taylor Van Vleet, Ullas Gargi, Sujoy Gupta, Yu He, Mike Lambert, Blake Livingston, and Dasarathi Sampath. 2010. The YouTube video recommendation system. In *Proceedings of the fourth ACM conference on Recommender systems* (RecSys '10). ACM, New York, NY, USA, 293-296. DOI=10.1145/1864708.1864770
[17] Bennet Ed. The Top 20 Most Popular Hospitals on YouTube. http://ebennett.org/top20y/
[18] Natalya F. Noy, Nigam H. Shah, Patricia L. Whetzel, Benjamin Dai, Michael Dorf, Nicholas Griffith, Clement Jonquet, Daniel L. Rubin, Margaret-Anne Storey, Christopher G. Chute, and Mark A. Musen. BioPortal: ontologies and integrated data resources at the click of a mouse. Nucl. Acids Res. (2009) 37 (suppl 2): W170-W173. http://nar.oxfordjournals.org/content/37/suppl_2/W170.short
[19] W. John Wilbur and Karl Sirotkin. 1992. The automatic identification of stop words. *J. Inf. Sci.* 18, 1 (January 1992), 45-55. DOI=10.1177/016555159201800106
[20] S Th Konstantinidis, L Fernandez-Luque, P D Bamidis, R Karlsen. The role of Taxonomies in Social Media and the Semantic Web for Health Education: A study of SNOMED CT terms in YouTube Health Video Tags. Methods of Information in Medicine. 2013 Feb: (52):2
[21] International Health Terminology Standards Development Organisation. SNOMED Clinical Terms User Guide - July 2008. International Release. International Health Terminology Standards Development Organisation. 2008.
[22] BioPortal REST services Wiki. [cited 2012 Jun 7]. Available from: http://www.bioontology.org/wiki/index.php/BioPortal_REST_services.