Kerstin Denecke, Peter Dolog, Pavel Smrz, Wolfgang Nejdl, Avaré Stewart (eds.)

Using Web Data in the Medical Domain

Proceedings of the
First International Workshop on Web Science and Information Exchange in the Medical Web, MedEx 2010
Raleigh, NC, USA, April 26, 2010

Sponsor:
M-Eco, http://www.meco-project.eu

Preface

The advent of Medicine 2.0 is increasingly making the Web a more accepted source of information for the medical domain and is also exploited for discussing medical problems and treatments. Health organizations monitor online news repositories and web pages for relevant data on epidemiological events. Physicians learn about the experiences of their colleagues provided through social media platforms such as weblogs, or forums. Moreover, patients can not only search for information, but also provide information about their experiences. This workshop is devoted to the technologies for dealing with social- and multi media for medical information gathering and exchange. It provides therefore an extension towards Web Science in general by focusing on one specific application domain, which is medicine. This area is very relevant for current research as well as the research community, government and industry.

Information gathering from medical social- and multimedia poses many challenges given the increasing content on the Web and the trade off of filtering noise at the cost of losing information which is potentially relevant. These issues are compounded by their impact on both information producers and consumers in the health care community.

This workshop is intended to intensify the exchange of ideas between various research communities involved in aspects related to the problem of accessing, exchanging, processing, filtering and making applications that rely upon health related Web information more reliable and adaptable. The submitted contributions published in these proceedings therefore reflect current research in this area: The topics range from content classification for Epidemic Intelligence and recommender systems for medical events to finding connections between texts or persons.

We would like to thank all members of the program committee for supporting us in the reviewing process, the organizers of the main conference WWW 2010 to which this workshop was co-located. We also would like to thank the authors for their willingness to revise their initial submissions based on the reviewers comments. Finally we would like to thank our invited speaker, Prof. Wendy Hall for her willingness to give a talk at our workshop.

April 2010

Kerstin Denecke and Peter Dolog
MedEx 2010 Program Chairs

Organization

Organising Committee

Kerstin Denecke	L3S Research Center, Hannover, Germany
Peter Dolog	Aalborg University, Denmark
Pavel Smrz	Brno University of Technology, Czech Republic
Wolfgang Nejdl	L3S Research Center and Knowledge Based System Department, Leibniz University Hannover, Germany
Avaré Stewart	L3S Research Center, Hannover, Germany

Program Committee

Ralph Grisham	New York University, USA
Clement Jonquet	Stanford University, USA
Richi Nayak	Queensland University of Technology, Australia
Natasha Noy	Stanford University, USA
Nigam Shah	Stanford University, USA
Pavel Smrz	Brno University of Technology, Czech Republic
Jim Warren	University of Auckland, New Zealand

Program Committee

International World Wide Web Conference (WWW)

M-Eco – Personalized Event-based Surveillance

European RTD context M-Eco is an EU-funded project that contributes to the area of Epidemic Intelligence. M-Eco will develop technologies for early detecting potential health threats in informal, textual information.

Short Introduction Many factors in today's changing societies contribute towards the continuous emergence of infectious diseases. Demographic change, globalization, bioterrorism, compounded with the resilient nature of viruses and diseases such as SARS and avian influenza have raised awareness for European society's increasing vulnerability. Traditional Epidemic Intelligence systems are designed to identify potential health threats, and rely upon data transmissions from laboratories or hospitals. They can be used to recognise long-term trends, but are limited in several ways. Threats, such as SARS, can go unrecognised since the signals indicating its existence may originate from sources other than the traditional ones. Second, a critical strategy for circumventing devastating public health events is early detection and early response. Conflictingly, the time with which information propagates through the traditional channels, can undermine time-sensitive strategies. Finally, traditional systems are well suited for recognising indicators for known diseases, but are not well designed for detecting those that are emerging. Faced with these limitations, traditional systems need to be complemented with additional approaches which are better targeted for the early detection of emerging threats. The Medical EcoSystem (M-Eco) project, will address these limitations by using Open Access Media and User Generated Content, as unofficial information sources for Epidemic Intelligence. This type of content has transformed the manner in which information propagates across the globe. Based on this, M-Eco will develop an Event-Based Epidemic Intelligence System which integrates unofficial and traditional sources for the early detection of emerging health threats. M-Eco will emphasize adaptivity and personalized filtering so that relevant signals can be detected for targeting the needs of public health officials who have to synthesize facts, assess risks and react to public health threats.

M-Eco Consortium The coordinator of M-Eco is the L3S Research Center, with the partners Aalborg University, Brno University of Technology, SAIL Labs Technology, Robert Koch Institut, Governmental Institute of Public Health of Lower Saxony, and Joint Research Center. Additional user institutions are participating in the project through the M-Eco Advisory Board, including representatives of the World Health Organiziation (WHO), European Center of Disease Control (ECDC), Health Protection Agency (HPA), Institut de Veille Sanitaire (INVS), and the Mekong Basin Disease Surveillance (MBDS).

Further Information http://meco-project.eu

M-Eco

Coordinator: Dr. Kerstin Denecke
(L3S Research Center, DE)

Table of Contents

Session 1

Session 2

Determining Patient Similarity in Medical Social Networks

Sebastian Klenk, Jürgen Dippon, Peter Fritz, and Gunther Heidemann

Stuttgart University
Intelligent Systems Group
Universitätsstrasse 38, 70569 Stuttgart, Germany
ais@vis.uni-stuttgart.de

Abstract. In social networks the primary concern of people is to find others who share similar interests. For medical systems this means finding people who have similar symptoms or comparable diseases. Here a simple matching of variables would lead to a very small number of identical cases and determining similarity would usually fail due to the categorical nature of most factors. In particular, such problems arise for cancer patients. We have developed a system that is capable of determining similarity in terms of the survival time distribution. By a similarity based search our approach allows to determine related patients. Thus recommendations for contacts of interest become possible. We will present the theoretical foundation as well as a use case scenario with an existing data mining software.

1 Introduction

Finding "patients like me" is a big issue for people suffering from severe illness. Today, this problem is addressed by the medical social network with identical name [1], and by organizations such as the german ACHSE[2] or the european Eurordis[3], which represent the common interests of patients and have brought together people with similar diseases successfully for several years now.

The goal of most medical social web sites is to provide a forum and a more direct way for patients to exchange thoughts, feelings, and experiences. Therefore the search for other people with a similar disease history and similar symptoms is crucial. For this purpose, patient profiles are presented which share a large number of similarities, just like in other social networks. However, defining such similarities for patient profiles is significantly more difficult than for other types of social networks. Different aspects of a disease have to be weighted differently, so a simple matching of factors is insufficient.

We have developed a similarity measure for cancer patients which calculates influence values for factor levels and thereby facilitates a soft matching. This

[1] PatientsLikeMe is a social networking health site with over 40,000 Members http://www.patientslikeme.com
[2] The German Alliance for Rare Chronic Diseases http://www.achse-online.de
[3] Eurordis – Rare Diseases Europe http://www.eurordis.org

means that different aspects are also weighted differently. For example, the fact that two cancer patients have developed metastasis is weighted much higher than similar age. This leads to a domain specific matching and provides better recommendations on who might have had similar experiences or who might have knowledge a user can benefit from. Finding relationships of this kind is the very basis of social media.

2 Related work

An important part of social networking research [17] is on *recommender systems* [5, 10, 8, 16]. These are systems that recommend certain items to the user, usually products, but also people one might want to know. As this is particularly interesting for e-commerce applications, most research is on suggesting new products.

For recommending people, there are two common approaches: (*i*) Content based recommendation which uses the information the user enters into the social network application, whereas (*ii*) relationship based recommendation traces who are the friends of the users friends, which the user might want to meet. Chen et. al. [2] provide an overview on both fields and perform a comparative study. Their results are mostly in favor of the relationship based approach, whereas they argue that similarity in content is so far calculated by keyword matching, which is just not sufficient. An example for a relationship based method is the work of Lin et. al. [12, 7], who deal with the problem of matching people in the context of searching for experts. They combine a graph based approach with a matching of search terms with profile terms which yields good results. But in the case of medical data an approach of this kind would lead to insignificant results because term matching does not reflect the true difference of the underling objects. Here more detailed domain knowledge is be required to determine term weightings. As stated by both Felfering et. al. [8] and Volinsky [16], deep domain knowledge is so far not used excessively in recommender systems.

It is obvious that content based recommendation is, at least in principle, superior to relationship based recommendation, as it would allow to explore the entire network rather than just the subset a user is connected to. We therefore aim at improving content based recommendation by making an interpretation of the given content feasible.

Another important aspect of a weighted content based approach is security. Such a system is less likely to be subject to fraud or spamming as described by Mobasher et. al [13].

Apart from recommender systems, distance learning has a long history in the area of case based reasoning [14]. Learning distance measures facilitates a context sensitive estimation of similar cases. Arshadi and Jurisica [1] employ logistic regression to estimate a distance measure which gives relevance to certain aspects of the data. The method we describe here differs from the one proposed by Arshadi and Jurisica, as it allows for continuous dependent data which can even be censored, a feature that is crucial for medical data.

The distance measure we are using here is based on an idea proposed in [6], which has been extended and implemented in the medical data mining system OCDM [11]. In the present paper we present a new application of this idea in the context of medical social networks.

3 Similarity for patient data

Measuring the similarity of natural continuous data items is very much straight forward. Every data dimension has the same weight and differences between dimensions can be interpreted in a very intuitive way. For categorical and artificial data, as is the case for patient data, differences in variables are anything else but intuitive and the weighting varies with each dimension. Formally speaking for two data items x and y a distance looks as follows:

$$d(x,y) = \sum_{k=1}^{n} \alpha_k d_k(x_k, y_k). \tag{1}$$

Here $\alpha = (\alpha_i)_{i=1\ldots n}$ is a weighting term that is assigned to each dimension and corresponds to its influence on the similarity. When working with lung cancer patient data for example it makes a huge difference whether the patient smokes or not but the area he or she lives in is of minor importance. Therefore similarities for smoker (yes or no) should have higher α values than for similarity in zip code.

Besides the weighting factor there is also the functions d_k which could be the absolute, the squared or the binary distance

$$d_k(x,y) = 1 \quad \text{if} \quad x = y \quad \text{else} \quad d(x,y) = 0$$

depending on the dimension k.

Determining a suitable weighting is essential to finding a good similarity measure. Therefore it is necessary to have a method at hands to calculate such a weighting. The central idea to the similarity measure learning approach we have taken, is to have a linear relation between a number of independent and one dependent variable that can be estimated and used as a weighting scheme. An ideal candidate to estimate such a scheme is the logistic regression [9]. This is a supervised learning scheme that, based on training data, estimates the influence a given set of independent variables has on a dependent variable.

Formally it calculates the probability of a variable G having a certain value g given the information contained in all the other variables $X = x$

$$P(G = g|X = x) = \frac{\exp(\beta_g^T x)}{1 + \sum\limits_{g' \in G} exp(\beta_{g'}^T x)}. \tag{2}$$

This formula gives us the influence each element x_i of x has on the outcome g of G. Here the weight vector β represents this information. Equation (2) can be used to model this influence for discrete data, for continuous and censored

dependent variables, Cox has developed a method to calculate β [3]. The central thought of his work is that the function $h(t|x)$ can be described as

$$h(t|x) = h_0(t) \cdot P(h = h_0|X = x), \tag{3}$$

where $h_0(t)$ is unknown. This leads to

$$h_0(t) \cdot \exp(\beta^T x). \tag{4}$$

What is actually estimated in (3) and (4) is the distribution function of the survival times. It is based on an unknown baseline hazard function that determines the risk of a patient at a certain moment. The formula in (3) is known as Cox proportional hazard regression or just Cox regression and is mostly used in survival analysis [4, 15]. The actual estimation of these parameters takes place with a Newton-Raphson based method. Therefore the partial log likelihood function (for the parameter β over a training set) is maximized.

Given the influence information β out of (3), it is easy to develop a distance measure that is sensitive to the relevant aspects of the data concerning the variable for which the estimator was trained.

Fig. 1. The recommendation of people in other social networks (on the left side Xing and on the right side Facebook)

3.1 Patient recommendations by regression estimation

Recommendations of other people in a social network is a central theme of social applications (see also Figure 3 for examples). In the above section we have described how regression estimation can lead to a weighting of variables and thereby allow for the calculation of specific distance measures. Here we will describe how such a measure can be used to determine other people in the social network that one might want to know.

A social network application consist of a large database containing information on the people belonging to it. The information was entered by the people

themselfs and may therefor contain only certain aspects of their profile. To determine other people with similar views it is necessary to calculate a distance measure as described above. Given a database with sample cases (the training data) one is able to estimate the weighting parameter β and apply it to a distance measure of the form:

$$d(x, y) = \sum_{k=1}^{n} \alpha_k d_k(x_k, y_k).$$

Here $\alpha = (\alpha_i)_{i=1\ldots n}$ with $\alpha_i = \exp^{\sigma \cdot \beta}$ and σ being a scaling factor to match the influence of the weighting on the distance measure. The measure itself could be the squared distance $d_k(x, \tilde{x}) = ||x - \tilde{x}||^2$ or simply the absolute distance. The scaling factor it self can be chosen to suite the needs of the recommendation, should the influence of the independent variables on the survival be weighted more heavily a value of $\sigma >> 1$ should be selected, in any other case $\sigma \leq 1$ is a good choice.

Now if this measure is applied to all people in the database one obtains a partially ordered list where the first few profiles can be used as recommendations. To reduced computational load one could restrict the number of computations by only considering profiles that share a least amount of common fields.

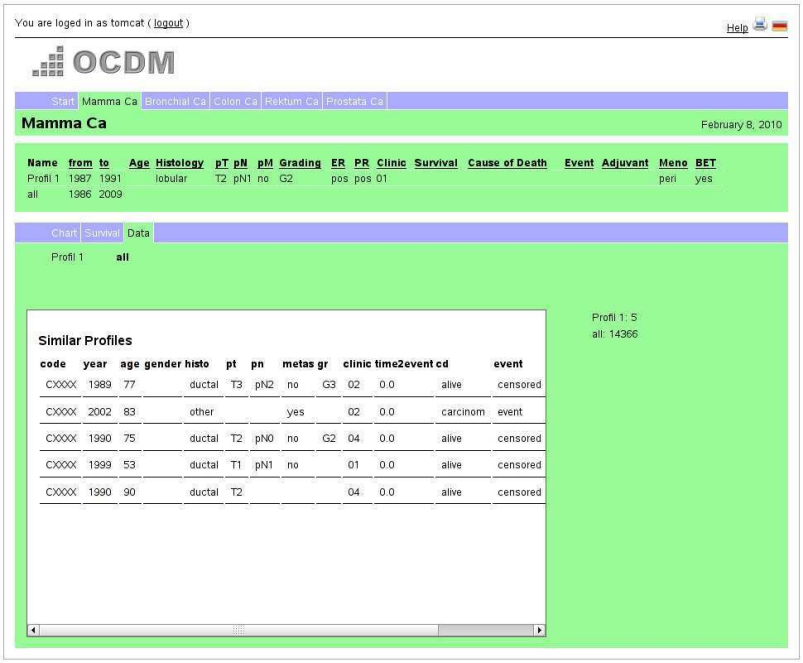


Fig. 2. The presentation of similar patients in the OCDM system

4 Implementation

We have implemented the similarity distance measure in our data mining software OCDM [11], where similar patients are found for a given patient profile. This system, although intended for physicians, recommends similar profiles for a further study. For a patient an identical approach could lead to a recommender system as described above. In this section we are describing technical details about the similarity search. We will thereby concentrate on rather generic technical aspects, further details about the actual implementation of the similarity search can be found in [11]. As basis for the developed system serves a PostgreSQL Database Server and a Java-Tomcat Servlet-Engine. As performance is a critical aspect of the software and much calculation has to be done during the estimation of the distance measure (on one hand the calculation of the weights and on the other the similarity calculation when recommending other profiles) we didn't follow a strict layer separation. Some tasks that involved extensive data processing were developed as stored procedures that run inside the database process. Most of the heavy-load calculation was thereby separated from the middleware and the GUI. As some of the calculation procedures are needed in the stored procedures and in the business logic we implemented these as Java classes such that they could be used in PL/Java code in the database as well as plain Java objects in the application server. We did some experiments with the similarity based distance we have developed and thereby achieved results comparable to that of common SQL queries. We measured the time it took for the database server to return results. For a data set of roughly 15.000 cases the database returned the select data on average after 10 milliseconds whereas the similarity based search took 35 milliseconds. These results can be placed in context when looking at the time it takes to process a simple SELECT statement with a function term (adding a constant to a column value) or a SELECT statement with an aggregate (calculating the average of a column value). The results are summarized in Table 1.

Query Type	mean	std-err.
Simple SELECT	9.29	6.28
SELECT with function term	24.81	8.29
SELECT with aggregate	95.60	11.31
Similarity Search	35.09	10.50

Table 1. Time until results are returned in milliseconds

5 Discussion

We have presented a domain specific distance measure for medical social networks. It is not intended to be generally applicable to the broad audience of

medical social networks, rather, it allows certain groups of patients to obtain better recommendations. If it is known that a user suffers, e.g., from a certain cancer type, search for other network members is focused and directed by criteria specific to this disease. The weighting in the actually calculated distance measure (1) can be easily adapted to a particular user group. Another important aspect of the above described distance measure is that it is solely focused on the survival time and does not include other possibly relevant aspects such as regional proximity or corresponding interests. In our experience, this restriction led to the best results. However, the restriction can be easily removed to include combinations of different weighting schemes. For two given weighting vectors α^1 and α^2 it is easy to combine them to a new weighting scheme α^* by just summing up corresponding normalized elements

$$\alpha_i^* = \frac{1}{2 \cdot ||\alpha^1||}\alpha_i^1 + \frac{1}{2 \cdot ||\alpha^2||}\alpha_i^2.$$

Data coding and treatment of missing values are important issues, because not every user will conform to standardized nomenclature to describe his or her disease, and likewise, many users will not present all their information in a social network. Both data coding and missing values have significant influence on distance estimation. Missing values can already be handled by the parameter estimation procedure and the distance measure itself as well. So the remaining problem is the lack of a formal notation. This, of course, could dramatically decrease the efficiency of the training process (if it is based on the data in the network). However, social networks have grown at such pace in the recent years that it is still highly likely to find a sufficient number of "good" training samples, even if data with unclear values have to be omitted. When it comes to proximity calculation, informal and varying notation could be handled in such a way that only those variable values that match certain criteria are considered for calculation, while all others are treated as missing values.

6 Conclusion

We have presented a method to calculate similarities of patient profiles for recommending people to other members in a social network. As connecting to other people is the central aspect of medical social networks, a subject specific similarity search can increase the performance of recommendations and thereby increase the usefulness of the social network application dramatically. In addition to presenting the theoretical foundation we also have given insight into some implementation details as well as performance measures. These show comparable results to more complex SQL queries and can serve as a guideline when implementing a similar approach in a real world application. As the method we have presented is highly subject specific, i.e., dependent on the estimation of survival time data, it might be interesting to see further research on other medical data that might be less dependent on a time to event. Further the incorporation of social graph information seems to be promising.

References

1. N. Arshadi and I. Jurisica. Data mining for case-based reasoning in high-dimensional biological domains. *Knowledge and Data Engineering, IEEE Transactions on*, 17(8):1127–1137, Aug. 2005.
2. Jilin Chen, Werner Geyer, Casey Dugan, Michael Muller, and Ido Guy. Make new friends, but keep the old: recommending people on social networking sites. In *CHI '09: Proceedings of the 27th international conference on Human factors in computing systems*, pages 201–210, New York, NY, USA, 2009. ACM.
3. D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society Series B (Methodological)*, 34(3):187–220, 1972.
4. David R. Cox and E. J. Snell. *Analysis of binary data*. Monographs on statistics and applied probability ; 32. Chapman and Hall, London, 2. ed. edition, 1989.
5. M. Deshpande and G. Karypis. Item-Based Top-N Recommendation Algorithms. *ACM Transactions on Information Systems*, 22(1):143–177, January 2004.
6. J. Dippon, P. Fritz, and M. Kohler. A statistical approach to case based reasoning, with application to breast cancer data. *Comput. Stat. Data Anal.*, 40(3):579–602, 2002.
7. Kate Ehrlich, Ching-Yung Lin, and Vicky Griffiths-Fisher. Searching for experts in the enterprise: combining text and social network analysis. In *GROUP '07: Proceedings of the 2007 international ACM conference on Supporting group work*, pages 117–126, New York, NY, USA, 2007. ACM.
8. Alexander Felfernig, Gerhard Friedrich, and Lars Schmidt-Thieme. Guest editors' introduction: Recommender systems. *IEEE Intelligent Systems*, 22:18–21, 2007.
9. Trevor J. Hastie, Robert J. Tibshirani, and Jerome H. Friedman. *The elements of statistical learning*. Springer, corrected print. edition, 2002.
10. Przemysław Kazienko and Katarzyna Musiał. Recommendation framework for online social networks. In *Advances in Web Intelligence and Data Mining*, Studies in Computational Intelligence, chapter 12, pages 111–120. 2006.
11. S. Klenk, J. Dippon, P. Fritz, and G. Heidemann. Interactive survival analysis with the ocdm system: From development to application. *Information Systems Frontiers*, 2009.
12. Ching-Yung Lin, Kate Ehrlich, Vicky Griffiths-Fisher, and Christopher Desforges. Smallblue: People mining for expertise search. *IEEE MultiMedia*, 15(1):78–84, 2008.
13. Bamshad Mobasher, Robin Burke, Runa Bhaumik, and Chad Williams. Toward trustworthy recommender systems: An analysis of attack models and algorithm robustness. *ACM Trans. Internet Technol.*, 7(4):23, 2007.
14. Petra Perner, editor. *Case-based reasoning on images and signals : with 30 tables*. Studies in computational intelligence ; 73. Springer, Berlin, 2008.
15. Steve Selvin. *Modern applied biostatistical methods using S-Plus*. Monographs in epidemiology and biostatistics ; 28. Oxford University Press, New York, 1998.
16. Chris Volinsky. Matrix factorization techniques for recommender systems. volume 42, pages 30–37, 2009.
17. A.C. Weaver and B.B. Morrison. Social networking. *Computer*, 41(2):97–100, Feb. 2008.

Tag and Neighbour Based Recommender System for Medical Events

Karunakar Reddy Bayyapu and Peter Dolog

IWIS — Intelligent Web and Information Systems,
Aalborg University, Computer Science Department
Selma Lagerlöfs Vej 300 DK-9220 Aalborg, Denmark
E-mail: {kreddy, dolog}@cs.aau.dk

Abstract. This paper presents an extension of a multifactor recommendation approach based on user tagging with term neighbours. Neighbours of words in tag vectors and documents provide for hitting larger set of documents and not only those matching with direct tag vectors or content of the documents. Tag popularity, tag representativeness and tag similarity are applied similarly as in the original approach but also to neighbours. By doing so, we treat the documents which have been added to the result set by considering word neighbours in the same way as the others. This provides an advantage in the situations where the quality of tags is lower. We discuss the approach on the examples from the existing Medworm system to indicate the usefulness of the approach.

1 Introduction

Search systems typically return a ranked list of web pages on different aspects of the same topic in the returned list in a response to a users request. Recently, social activities such as tagging have emerged mostly to help people to organize resources of their personal interest on the web. The tagging information has been applied to help information retrieval and recommender systems. Although some successful applications have been developed (see, for instance,[4]), implementing and extending a hybrid tag-based recommender system with personalization for social bookmarking systems is still a challenge. Many applications would benefit from tag based analysis, which is sufficiently general to be useful in a wide range of applications, is already performed.

Tags in Medical bookmarking systems such as $Medworm$ are usually assigned to organize and share resources on the Web. Tag clouds are weighted lists of tags. The relative importance of a tag is visualized with bigger font size, bolder letters, and is measured as a count of the popularity of the tag i.e. how many times users have used it to describe a resource. Tags are generally submitted by any user in bookmarking services such as for example $http://medworm.com$. The data source tags provide useful information with sufficient background and context, even though the set of tags are quite limited to provide an accurate degree of relatedness between tags and neighbours.

Furthermore, the tags themselves as they appear in the*Medworn* system, for example, represent multiple domains. Therefore, it is not directly applicable to consider only tag measures. In our approach, we focus our efforts on neighbour objects of the tags for search and retrieval systems. Our approach combines basic similarity calculus with external factors such as a tag popularity, tag representativeness and closest neighbours semantic similarity, document score, semantic similarity of tags as described in [4]. The proposed contribution of this paper is:

- The extended hybrid tag based recommender system which bases the computation on a user query,
- Finding the closest neighbour vector of the tag from medical data source.

The rest of the paper is structured as follows. Section 2 discusses the problem and proposed solution on an example. Section 3 defines our approach to multi-factor recommendation with word neighbours. Section 4 discusses related work and positions our work in this context. Section 5 discusses experimental evaluations and outcome results. Section 6 explains analysis of experiment and conclusions .Section 7 discusses future work.

2 Working example and Motivating scenario

Let's consider the following scenario. If the user submits a query to the system, the system evaluates which documents are relevant to the query and returns a rank ordered list of documents to the users. Normally, the system considers content of the documents or collaborative user activities as factors to judge the relevance. Recently, tag-based approaches have been coined in the literature as well but only in the general situations. The domain specific searches by experts usually do not hit the most relevant documents in the first top n results. For example, the medical *Medworm* system gives almost 14470 records just based on swine flu tag. However, the system does not know what exactly user is looking for, or user doesn't know the proper words to describe what it is that he wants. Then the returned results are often unsatisfactory.

However, we could resolve this problem by proposed extension of tag and neighbour based recommender systems for medical events. This approach searches for the most trusted information. The traditional approach based on simple tag based recommendation factors are not efficient enough for domain specific systems such as *Medworm* due to lower relevance of user generated tags used. Therefore, we also apply neighbours to consider wider set of documents.

Figure 1 shows which kind of information *Medworm* can offer to improve the search to find information with respect to a certain aspect of a document. One just needs to refer to its associated tags and predicted neighbours in the corresponding documents. There are two columns depicted in the figure which represent the positions of neighbours in the text or the tag vector. By considering one side or both sides of the neighbourhood, we can target wider space of documents than with original query. This allows for expansion of the document set hit by the query. Therefore, we ensure that the user will not miss important

medical events documented in a document or a blog even when it does not match exactly the query. By applying personalization factors as in original query to extracted neighbours, we achieve similar ranking and therefore provide a means to access the most relevant documents.

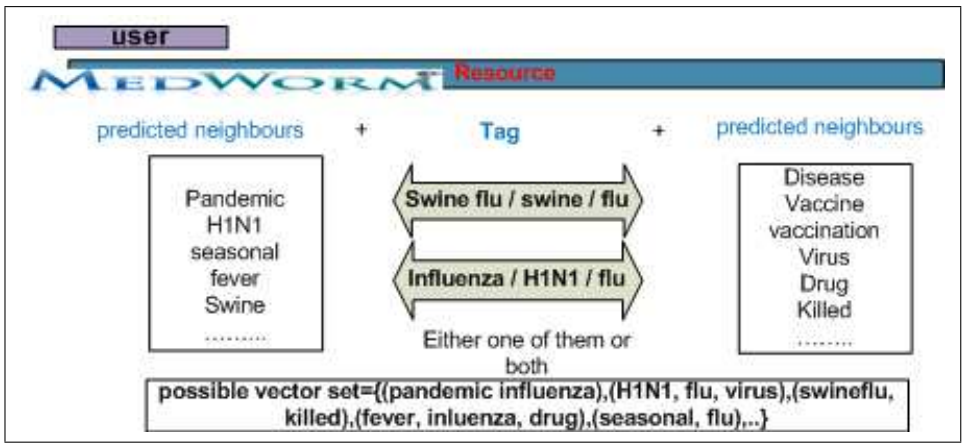


Fig. 1. Relation between predicted tags and their neighbours of user?s interest

3 Recommender System

3.1 Concept

The main concept is to achieve recommender system in medical events. The proposed recommender system combines associated neighbours with different aspects of similarity and tags. Consequently, a hybrid recommender system obtained integrate many independent recommendations by applying tag popularity, tag representativeness, tag similarity and neighbours. It exploits neighbours that are dynamically re-calculated according to the effectiveness of the recommendation. The system also uses the semantic similarity between the neighbours to calculate neighbours set. These results influence the final recommendation list to re-order the rank. The process of the calculation looks as follows. First, a users click on a tag or he submits a query. Then, the recommendation algorithm is applied to produce a set of recommended resources. Second, this set is then sorted by taking the user interest and tag neighbours into account and re-ranks the results accordingly.

3.2 Multi-Factor Recommendation with Word Neighbours

The tag and neighbour based recommender approach based on [4] is calculated as:

The extended hybrid similarity score (HS),

$$HS_{(D_i,D_{ii})} = [(Ds_{D_i} \times Ds_{D_{ii}}) \times (TS_{(D_i,D_{ii})})] \times NS_{(D_i,D_{ii})} \ ,$$

where D_i and D_{ii} are a particular documents from a set of documents D. Ds is the document score, TS is a function for measuring the tag similarity and NS is the closest neighbour vector space. We define document score as [4]:

$$Ds = \sum_{i=1}^{n} Popularity(Tag_i) \times \sum_{i=1}^{n} Representativeness(Tag_i),$$

where n is the total number of existing tags in the repository and for the definitions see in [4]. Informally, each one of the factors in the above formulas is calculated as follows:

Tag Popularity. The tag popularity is calculated as a count of occurrences of one tag per total of resources available [4]. We rely on the fact that the most popular tags are like anchors to the most confident resources. As a consequence, it decreases the chance of dissatisfaction by the receivers of the recommendations.

Tag Representativeness. [4] It measures how much a tag can represent a document it belongs to. It is believed that those tags which most appear in the document can better represent it. The tag representativeness is measured by the term frequency.

Tag Similarity (TS). It combines the classical cosine similarity (CosSim) from user query and information retrieval field with a semantic similarity (SemSim) which is defined in [4].

Cosine similarity (CosSim). The cosine similarity in our approach is a measure between, a user query Q transformed into a tag vectore and a set of tags or words (represented as a vector) of particular web page ($\bar{W}$). Each word, w(ti), in each dimension corresponds to the importance of a particular tag ti.

The Cosine similarity (Q, wi) is calculated for every tag word resource $wi \in \bar{W}$. As an output, this stage of the algorithm will produce a subset of resources W', that have some similarity to the query tag and similarity scores for each [15].

Let us assume that the user interacts with the system by selecting a query tag and expects to receive resource recommendations. Therefore, a query is a unit vector consisting of a single tag, and the equation is (adapted from [12]):

$$CosSim(Q_{D_i,wD_{ii}}) = \frac{\bar{W}_{(Q_{D_i},w_{D_{ii}})}}{\sqrt{\sum_{t \in T} \bar{W}_{(Q_{Di},w_{D_{ii}})}}}$$

where T is the set of tags, the similarity of the selected tag to each resource and recommends the top n., $\bar{W}$ is the set of words of that particular visited web page, D_i and D_{ii} are a particular documents from a set of web pages/documents.

Semantic Similarity (SemSim). The semantic relation between two tags is defined as follows:

$$SemSim(s,t) = MDSim(s,t) \times OntoSim(s,t), \forall s,t \in T$$

Where s and t are particular tags of set of tags T. MD Sim (s,t) is the Medical Dictionary similarity score and OntoSim (s,t) is the similarity score achieved from ontologies.

Neighbour Semantic Similarity (NS). Calculates closest neighbour influence for personalized recommendation by vector space models [13]. In order to find the nearest neighbours of the tag word, it must measure the similarity of the tag words [20], and select several words that have the highest similarity as the nearest neighbours of the tag word. We adopt cosine similarity algorithm to measure the similarity between word $w(ti)$ and $w(tj)$. If the user does not rate the words, we can assume the rating is zero. Assuming the rating of the n-dimensional word space of word $w(ti)$ and $w(tj)$ is respectively vector $w(\bar{ti})$ and $w(\bar{tj})$.

The similarity between word $w(ti)$ and $w(tj)$ is $sim(w(ti), w(tj))$

$$sim(w(ti), w(tj)) = \cos w(\bar{ti}), w(\bar{tj}) = \frac{w(\bar{ti}) * w(\bar{tj})}{\|w(\bar{ti})\| * \|w(\bar{tj})\|}$$

In this view, the solution to be addressed includes how to represent the tags and their closest neighbours and how to use it to influence the activation of user preferences. The approach is to predicted neighbours, the current visited context is represented as (is approximated by) a set of words concepts from the domain ontology. Ultimately, the perceived effect of neighbours extended hybrid approach is that user interests that are in focus for a current context, and those that are in the semantic scope of the ongoing user activity are considered for personalization [3].

4 Related Work

Tags have been recently studied in the context of recommender systems due to various reasons. Tags are signals or labels that particular resource was interesting for a user, and he bookmarked it as well as tagged it with a specific tag relevant for a particular situation a user was encountered in. Recommendations of relevant events should be based on the sufficient occurrences for similar signals expressed by tags. Therefore, different similarity measures need to be studied in this context for effectiveness and efficiency. [10] argues for a solution where tagging from social bookmarking provides a context for recommender systems in terms of context clues from tags as well as connectivity among users to improve the collaborative recommender system. [11] constructed a web recommender based on large amount of public bookmark data on Social Bookmarking systems. For means of personalization, [11] utilizes folksonomy tags to classify web pages and to express user's preferences. By clustering folksonomy tags, they

can adjust the abstraction level of user's preferences to the appropriate level. [11] experiment did not measure the efficiency of the recommendations in terms of user satisfaction what could give us a parameter for comparison. [17] extends a content based recommender system by deriving current and general personal interests of users from different tags according to different time intervals. However, the similarity of the tags is given by two Na?ve Bayes classifiers trained over different timeframes: one classifier predicts the user's current interest, whereas the other classifier predicts the user's general interest in a bookmark. The two classifiers are trained with a subset of the bookmarks created by a user. The tags of each bookmark, converted into a "bag of words", are used as training features. The bookmarks are recommended in the case of both two classifiers predicting a bookmark as interesting. The effectiveness of the recommendations, however, is totally dependent on the quality of the subset of bookmarks used for training the classifiers.

[19] proposes a collaborative filtering approach TBCF (Tag-based Collaborative Filtering) based on the semantic distance among tags assigned by different users to improve the effectiveness of neighbour selection. That is, two users could be considered similar not only if they rated the items similarly, but also if they have similar understanding over these items. To calculate the semantic similarity, the WordNet dictionary is being accessed to find the shortest path connecting a tag and its synonym in the graph synsets. The semantic distance based calculation, which might be difficult depending on the context of users. Special vocabularies hardly are found in general purpose dictionaries such as WordNet. Furthermore, the WordNet lacks much data useful to support proper name disambiguation, and it is not collaboratively edited [8]. [7] develops a page rank based algorithm for recommendations of resources based on preference vectors in folksonomy systems. [5] shows the benefits of using tag based profiles for personalized recommendations of music on Last.fm. The purpose of tags varies as well as tagging itself may be influenced by different factors. For example, [14] studies a model for tagging evolution based on the community influence and personal tendency. It shows how 4 different options to display tags affect user's tagging behavior. [1] studies how the tags are used for search purposes. It confirms that the tags can represent a different purpose such as topic, self reference, and so on and that the distribution of usage between the purposes varies across the domains. It compares the purposes with another literature (such as [6, 18, 14]) where these are called differently.

Other works such as [16] and [9] coined the term emergent semantics as the semantics which emerge in communities as social agreement on tag's meaning that the semantics is derived from its frequent use instead of the contract given by ontologies from ontology engineering point of view. However, the approaches based on emergent semantics are characterized by the power law which gives a long tail of the tags of which semantics have not emerged yet. Therefore, [2] looks at grounding of the tag relatedness with a help of WordNet.

5 Evaluation and Results

We have conducted an experiment to preliminary assess the performance of the recommender approach proposed in this paper. The nature of the experiment was based on a simulation a mix of scenarios regarding the amount of pages, tags and their neighbours. The proposed scenarios were created aiming at simulating realistic usage of $Medworm$. The variables addressed by each scenario are:

- *Amount of Pages:* each page has a set of tags that are compared for processing the recommendations. Therefore, the more pages exist then more time will be spent to calculate the similarity between the pages.
- *Amount of Tags:* the similarity of the pages is given by their tags. The whole set of tags of each page must be compared to verify which ones are similar.
- *Set of neighbours:* set of neighbours are depending on similarity of the tags. The whole set of neighbours of each page must be compared with semantic meaning of the tag.

These variables were chosen because we are using them for calculating the recommendations. This process is time consuming and invariably affects the system performance. However, it does not mean that other factors such as page size should not be considered[4].

We found that the choice of $tf * idf$ played an important role to find tag representativeness. In our evaluation, $tfidf$ have identical trends, but $tfidf$ always provides superior results, so we have reported only results found based on those weights. We were able to extract tag neighbours. Some samples were taken from each dataset of neighbours to find neighbours semantic similarity using cosine similarity. The validation was performed to measure the improvement in recommendation. We used MedicineNet[1] online free medical dictionary to measure semantic similarity.

Let's assume that given semantic similarity is $\varphi(s)$, where S is the semantic similarity. We also have to define cosine similarity between the user's query and tags ti, tj, tk, tl, etc. These tags could be in vector node of current webpage: $\bar{W} = \{\ldots, ti, \ldots, tj, \ldots, tk, \ldots, tl \ldots\ldots\}$, where $\varphi(s) \subseteq \bar{w}$ and $\varphi(s) \cap \bar{w} = \emptyset$

The neighbour similarity depends on the similarity of the tag words. So, neighbour semantic similarity (NS) is subset of cosine similarity. NS can be computed as follows:
$\bar{\varphi}(NS) = \{\ldots, ti-1, ti+1, \ldots, tj-1, tj+1, \ldots, tk-1, tk+1, \ldots, tl-1, tl+1, \ldots\ldots\}$, $\bar{\varphi}(NS) \subseteq \bar{w}$ and $\bar{\varphi}(NS) \sim \varphi(s)$.

In order to test the effectiveness of the algorithm, we compute the factors with a collection of documents from $Medworm$ data source about 90 pages. The documents were encoded with xml format, so we decided to make 466 tags manually. Content of the pages and some tags were extracted from web sites on the Internet. Similarly, we utilized manually generated neighbours to assign tags to $Medworm$ pages tagged by a particular user. Due to certain constraints,

[1] http://www.medicinenet.com

we had to limit the number of user queries. We needed just adequate number of satisfactorily different pages and sufficiently different assignment of tags to them. Each test case consists of tag, neighbours, semantic factors and resource. We consider this resource as the target results, since we know that the user is interested in it.

Here, we have one user interest query "avian flu pandemic". A simple way to start out is by eliminating documents that do not contain all three tags "avian", "flu", "pandemic", but in general it hits many documents. Our algorithm distinguished relevant and irrelevant documents and tags like "avian", "flu", "pandemic" that occur rarely and good keywords. After performing a recommendation using both the tag and neighbours, the rank of the target resource in the recommendation set was recorded and shown in table1.

Table 1. First five documents with highest HS returned by our Hybrid Recommendation approach for a query= *avian flu pandemic*. The top document entitled *Birds in the news*, is intuitively relevant to the query

ReturnPos	Document#	Document score(DS)	Recommedation Score(HS)
1	34	19.12	14.531
2	12	14.06	8.857
3	56	12.03	4.879
4	8	10.05	3.601
5	44	10.79	2.421

As seen at Table 1 our extended hybrid based algorithm accepts a user interest, a set of neighbours, and a selected tag. The recommendation and document scores are from the interval between 0 and 20. As we can see, the top recommendation score is 14.53 (72.65% accuracy). It means that the particular document is the most relevant to users query. Result positions 2, 3, 4 are also relevant to user query but not most relevant when comparing to position 1. Position 5 document got mixed results and it is partially related to the query. This result was not considered excellent but satisfactory since our recommendations relied basically on syntax similarity of the tags.

6 Discussions and Conclusions

Categorized *Medworm* is the application of medical RSS feeds [2] to better target the delivery of health care, facilitate the discovery of new products, and helps to determine a person's predisposition to a particular disease or condition through web. In this recommender systems, the extended hybrid algorithm

[2] http://www.medworm.com/rss/aboutmedworm.php

can perform tasks such as discovering documents (much like the web robots), ranking documents, filtering them, and automatically routing useful and interesting information to users and it has learning and adaptation capabilities. In fact, the extended hybrid algorithm perfectly suits information discovery and retrieval in the web. For example, information discovery and ranking can be handled by document score which depends on the tag popularity and tag representativeness. Another tag similarity can specialize in indexing, yet another like neighbours similarity can implement an information retrieval, and so forth. Applying these techniques to the web pages/documents, retrieved by a search tool could substantially weed out unrelated documents and improve the ranking quality of the remaining page/documents [15].

As per our experiment, the user has tagged several medical object web resources with the tag "flu". If a user selects that tag, the system should recommend resources concerning the number of records about the "flu". Certainly in addition, another "flu" related documents may have been tagged with alternative tags: disease, swine, H1N1, avian, bird flu, influenza, flu attacks, respiratory problems, pandemic, symptoms, lung infection, etc. These resources may have been tagged with "flu" and may not have been "flu" but they should still be made available to a user. The recommendation strategies must also be adapted to deal with the neighbours. Typically, recommender systems have dealt with two dimensions: users query and object neighbours semantic similarity. Consider again the "flu" related topics. After selecting "flu" the system may recommend resources only related by semantic similarities of neighbours. User may notice as of his query "avian flu pandemic", the "avian", "flu", and "pandemic" are related tags. These are generated by the closest neighbour semantic objects.

Finally, user may notice resources in *Medworm's* profile dealing with the medical neighbour objects, and view one of those resources as his priority is most trusted information in the top position.

7 Future Work

The evaluation of the system provided some supplementary conclusions, namely, a recommendation performed with association neighbours appeared to be the most useful but only positive influence neighbours. Future work will focus on positive and negative neighbours and tag similarity i.e. sentiment analysis. It would benefit the system to utilize such opinions and to lower the score of bad results, even if other strategies show them as recommendable.

Acknowledgments

This work is partially supported by the European Union under the IST project M-Eco (http://www.meco-project.eu).

References

1. Kerstin Bischoff, Claudiu S. Firan, Wolfgang Nejdl, and Raluca Paiu. Can all tags be used for search? In James G. Shanahan et al., editor, *Proc. of the 17th ACM Conference on Information and Knowledge Management, CIKM 2008*, pages 193–202, Napa Valley, California, USA, October 2008. ACM.
2. Ciro Cattuto, Dominik Benz, Andreas Hotho, and Gerd Stumme. Semantic grounding of tag relatedness in social bookmarking systems. In Amit P. Sheth et al., editor, *The Semantic Web - ISWC 2008, Proc. of the 7th Intl. Semantic Web Conference, ISWC 2008*, volume 5318 of *Lecture Notes in Computer Science*, pages 615–631, Karlsruhe, Germany, October 2008. Springer.
3. C. Dolbear, P. Hobson, D. Vallet, M. Fernndez, I. Cantador, and P. Castells. Personalised multimedia summaries. In *Semantic Multimedia and Ontologies Part III*. 2008.
4. Frederico Duarao and Peter Dolog. Extending a hybrid tag-based recommender system with personalization. In *Accepted for ACM Symposium on Applied Computing 2010*, Lausane, Switzerland, 2010. ACM Press. Accepted for publication.
5. Claudiu S. Firan, Wolfgang Nejdl, and Raluca Paiu. The benefit of using tag-based profiles. In Virgílio A. F. Almeida and Ricardo A. Baeza-Yates, editors, *Fifth Latin American Web Congress (LA-Web 2007)*, pages 32–41, Santiago de Chile, November 2007. IEEE.
6. Scott A. Golder and Bernardo A. Huberman. Usage patterns of collaborative tagging systems. *Journal of Information Science*, 32(2):198–208, 2006.
7. Andreas Hotho, Robert Jäschke, Christoph Schmitz, and Gerd Stumme. Information retrieval in folksonomies: Search and ranking. In York Sure and John Domingue, editors, *The Semantic Web: Research and Applications, Proc. of the 3rd European Semantic Web Conference, ESWC 2006*, volume 4011 of *Lecture Notes in Computer Science*, pages 411–426, Budva, Montenegro, June 2006. Springer.
8. Jay J. Jiang and David W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. Technical report, 1997.
9. Peter Mika. Ontologies are us: A unified model of social networks and semantics. *Web Semant.*, 5(1):5–15, 2007.
10. Reyn Nakamoto, Shinsuke Nakajima, Jun Miyazaki, and Shunsuke Uemura. Tag-based contextual collaborative filtering. *IAENG Intl. Journal of Computer Science*, 34(2), 2007.
11. Satoshi Niwa, Takuo Doi, and Shinichi Honiden. Web page recommender system based on folksonomy mining for itng '06 submissions. In *ITNG '06: Proc. of the Third Intl. Conference on Information Technology: New Generations*, pages 388–393, Washington, DC, USA, 2006. IEEE.
12. C. Papakonstantinou, I. Panagiotou, and F. Verbeek. Tag based meta-search for browsing the web: The tictag application. In *Proceedings 13th Computer-Human Interaction Netherlands Conference*, Leiden, The Netherlans, 2009.
13. G. Salton, A.Wong, and C. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.
14. Shilad Sen, Shyong K. Lam, Al Mamunur Rashid, Dan Cosley, Dan Frankowski, Jeremy Osterhouse, F. Maxwell Harper, and John Riedl. tagging, communities, vocabulary, evolution. In Pamela J. Hinds and David Martin, editors, *Proc. of the 2006 ACM Conference on Computer Supported Cooperative Work, CSCW 2006*, pages 181–190, Banff, Canada, November 2006. ACM.

15. A. Shepitsen, J. Gemmell, B. Mobasher, and R. Burke. Personalized recommendation in social tagging systems using hierarchical clustering. In *Proceedings of the 2008 ACM conference on Recommender systems*, pages 259 – 266, 2008.
16. Steffen Staab. Emergent semantics. *IEEE Intelligent Systems*, 17(1):78–86, 2002.
17. Pavan Kumar Vatturi, Werner Geyer, Casey Dugan, Michael Muller, and Beth Brownholtz. Tag-based filtering for personalized bookmark recommendations. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge mining*, pages 1395–1396, New York, NY, USA, 2008. ACM.
18. Zhichen Xu, Yun Fu, Jianchang Mao, and Difu Su. Towards the semantic web: Collaborative tag suggestions. In *WWW2006: Proc. of the Collaborative Web Tagging Workshop*, Edinburgh, Scotland, 2006.
19. Shiwan Zhao, Nan Du, Andreas Nauerz, Xiatian Zhang, Quan Yuan, and Rongyao Fu. Improved recommendation based on collaborative tagging behaviors. In *IUI '08: Proc. of the 13th Intl. conference on Intelligent user interfaces*, pages 413–416, New York, NY, USA, 2008. ACM.
20. L. Zheng, Y. Wang, J. Qi, and D. Liu. Research and improvement of personalized recommendation algorithm based on collaborative filtering. *IJCSNS International Journal of Computer Science and Network Security*, 7(7), jul 2007.

Can ProMED-mail Bootstrap Blogs? Automatic Labeling of Victim-reporting Sentences*

Avaré Stewart and Kerstin Denecke

L3S Research Center
Appelstr. 9A, 30169 Hannover, Germany

Abstract. Due to the proliferation of social media data and user-generated content available, monitoring trends or using this data in other scenarios becomes more interesting. Our research focuses on the extraction of information on health events from user generated content with the objective to support Epidemic Intelligence. Specifically, we describe and evaluate a method for identifying sentences relevant for event extraction. Labeled data is unavailable for this task and manual annotation is expensive. Therefore, in order to reduce the number of labeled examples, we apply a bootstrapping algorithm for this task. In more detail, we will study the suitability of a classifier trained on one text type (e-mails) for the classification of texts of another text type (blogs).

1 Introduction

The spread of infectious diseases and - due to this - the increased public concern - raises the necessity to have health surveillance systems on hand for detecting disease outbreaks as early as possible. All the activities related to early identification of potential health hazards, their verification, assessment and investigation with the objective to recommend public health control measures are summarized by the term *Epidemic Intelligence* [1]. A *public health event* is an event that creates a need for action of public health officials, for instance an outbreak of an infectious disease or one case of a very seldom infectious disease. A public health event can be described by *event information* providing information on *who* was infected by *what*, *where* and *when*, i.e., information on a victim, a location, a time or a disease.

Besides the traditional surveillance systems that monitor indicators such as death rates, drug prescriptions, occurrence of viruses etc. event-based systems have been developed. They extract and analyse outbreak-related information from text in electronic sources such as e-mail, official reports, news wires and present the results to the user. Social media data or user-generated content (e.g., Weblogs, Twitter messages) remained so far unconsidered for Epidemic Intelligence. In our research, we will focus on this text type.

The problem of detecting health events can be decomposed into mainly three sub-problems:

* This work has been done within the M-Eco project, partly funded by the European Commission under 247829.

1. Annotation: Identifying sentences containing information on an event.
2. Event Extraction: Identifying relevant facts to describe the event.
3. Event Aggregation: Aggregating information on the same event that has been reported in different sources.

In this paper, the focus is on identifying pieces of text relevant for public health event detection (Annotation problem). State of the art approaches for detecting public health events either rely upon a huge number of extraction patterns or a large set of labeled data in order to detect the relevant information-bearing sentences. However, when extracting such information specifically from social media, additional challenges are faced which make these approaches inadequate[2] including use of specific language, and different styles of writing. Large amounts of social media data are available. The data is noisy to a large extent; and it is often opinionated and can contain irrelevant information.

We address these challenges using semi-supervised learning in the context of event extraction. In particular, the sub-problem of identifying sentences relevant to the detection of health events on infectious diseases is considered. We will study the suitability of a classifier trained on one text type for the classification of texts of another. In more detail, patterns acquired from a manually curated source are transferred to bootstrap the classifier for medical blogs. To reduce manual work for labeling training data, we come up with an approach of automatically labeling a training set that bases upon research results known from the field of text summarization. The automatically determined training set is then used to learn a classifier for sentence classification.

The contributions of this work are (1) presentation of a health event detection framework, (2) introduction of a cross-corpus bootstrapping framework to identify relevant sentences, and (3) study related corpora for the task of bootstrapping.

2 Related Work

Our final goal is to extract public health events for outbreak detection. Existing systems for this task rely upon the enumeration of possible types of victim reporting patterns (e.g., MediSys [3], HealthMap [4], BioCaster [5]). MediSys for example uses manually specified keywords in different languages to identify news articles reporting on health events. In these systems, little or no attention is devoted to using blogs and other forms of user generated content as source of information. Other systems rely upon the linguistic, interpretive, and analytical expertise of analysts to filter and extract information about health threats (e.g., GPHIN [6]). Given the dynamic nature of social media in general and of weblogs in particular, a pattern-based system is not realizable since the number of patterns needed may be numerous. Further, the large amount of user generated content available requires application of automatic methods. In this paper, an approach based on semi-supervised learning for identifying relevant sentences for event extraction is introduced.

Bootstrapping approaches are such semi-supervised learning approaches and can be grouped into self-training and co-training approaches. Semi-supervised self-training has been applied in the field of sentiment analysis (e.g., subjectivity analysis [7]). Didaci and Role [8] compare several semi-supervised learning methods for multiple classifier systems. Chen and Ji [9] present a framework where bootstrapping is used for event extraction in a cross-lingual setting. An event extraction system in one language is bootstrapped by exploring new evidences from a system in another language. In this work, we apply an existing bootstrapping algorithm to the task of sentence classification. To the best of our knowledge, this particular task of classifying sentences of user generated content for health event detection has not been considered before. Further, bootstrapping has not been used in this context before. We will study the quality of such approach and report the results.

Methods which make use of unannotated text and intra-document information have emerged as important approaches for information gathering. The systems typically rely upon the redundancy present on the web, and assume that facts with multiple mentions are more reliable [3]. Zhen and Li [10] consider the problem of cross-domain text classification in the news domain. They propose a support vector based semi-supervised algorithm to solve this problem. Yangarber et al. [11] use cross-document analysis to support building a consistent and robust fact base about epidemics or the outbreak of disease. In our work, the cross-corpora analysis is applied to detect victim-reporting sentences. Cross-classification is applied to train a classifier on a data set of an auxiliary domain to classify data of the target domain. Besides reporting quality results of the approach, we will study the conditions when it is possible to use such a learner.

3 Approach

The objective of our approach is to identify information bearing sentences in social media, when faced with the problem of large amounts of unlabeled data. In particular, we consider a sentence relevant when it contains information on disease outbreaks (see section 4 for more details).

Given the characteristics of social media data, a supervised classification approach seems to be better suited than pattern-based approaches that rely upon extensive manual work. Our objective is to reduce the manual work for labeling examples as much as possible. For this reason, we make use of data of an auxiliary domain to determine labeled training examples. In more detail, for an auxiliary domain, we build a classifier in a bootstrap process. This classifier is then applied to label sentences of the target domain. To avoid manual labeling, we introduce an approach to automatic labeling examples of the auxiliary domain. The single processing steps are described in more detail in the next sections.

3.1 Automatic Labeling

For learning the classifier, a related and complementary, auxiliary data source is used. This related source typically uses a terse and more compact style of prose. It

also exhibits structural properties which allow us to apply a weak form of labeling - i.e., automatically label selected sentences as positive or negative examples with respect to disease reporting based on their position in the document. This is in contrast to the sentences in the target domain, for which we have a less obvious structural pattern from which we can weakly label sentences.

In more detail, the idea is to use sentences at the beginning of a document as positive examples. This idea has already been proven successful in the field of text summarization where sentences at the beginning of a document are used to produce a document summary.

Let $D = d_1, d_2, \ldots, d_j$ be the set of documents in an auxiliary corpus, where each document, $d_i \in D$, consists of a set of one or more sentences. Further, let $T = t_1, t_2, ..., t_m$ be a set of feature types used to represent the sentences of D. We kept the approach general with respect to the set of features to be used. Types of features can include bag-of-words, bag-of-concepts, part of speech, or a typed dependency structure.

Given a set of documents D and a surrogate representation for a given type, T_t applied to the sentences in D, a corpus can be modeled as a sentence database, $S_t = s_{t11}, s_{t12}, ..., s_{tjk}$, where s_{tjk} represents the k^{th} sentence for the j^{th} document using feature type, t. Further, we label the top-N sentences in the database automatically as positive cases and the bottom-N as negative examples, for a threshold value of N.

At this stage, we also want to introduce the concept of *sublanguages*. Given a sentence database of type t, S_t, for an auxiliary domain. Then, we can define the auxiliary corpus to be a sublanguage for the target corpus if a self-trained classifier built from the auxiliary domain performs well on the unlabeled examples in the target domain with some threshold tolerance. In our experiments, we will study whether the dataset from the auxiliary domain is a sublanguage for the target corpus.

3.2 Cross-Corpora Bootstrapping

Using the previously described approach to automatically labeling the sentences of the auxiliary domain, we determine a set of labeled examples to be used to train a classification model using bootstrapping. We are considering these examples produced by the automatic labelling approach as *weakly labeled*, i.e., there is some confidence why they have been selected as positive or negative examples, but it is unclear to what extent this labeling is confident. To reduce bias produced by this incertitude, we produce an improved classifier through a bootstrap process. The bootstrap process is depicted in Figure 2. The algorithm is described in more detail in the following (see Figure 1).

3.3 Classifying Sentences

The previous step produces a classifier trained on material of the auxiliary domain. We have chosen Support Vector Machines as classification algorithm. Finally, this classifier is applied to the sentences of the target domain and labels

Input:
W: a set of weakly labeled examples from an auxiliary domain
L: set of labeled examples from the auxiliary domain
C_i: a set of classifiers

Initialization:
Initialize the set L by selecting the weakly labeled candidates from W. Create a pool W_{prime} of examples by choosing p examples from W

Loop until a condition is satisfied (e.g., W = $\emptyset$, or threshold number of iterations reached)

- Train each classifier C_i on L, and label the examples in W_{prime}
- For each classifier C_i, select the most confidently labeled examples (e.g., the confidence score is above an absolute or relative threshold for the current run) and add them to L
- Refill W_{prime} with examples from W, and keep the size of W_{prime} as constant p

Fig. 1. Bootstapping Algorithm

the sentences of the target domain as positive or negative, or victim-related and not victim-related, respectively.

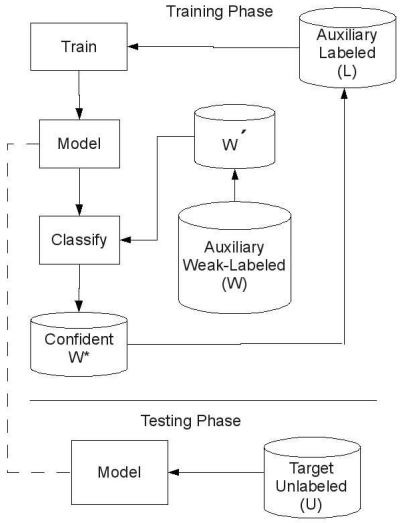


Fig. 2. Weakly Labeled Bootstrap: Overview

4 Experiments

In this preliminary work the experimental goals are twofold. First, we are interested in knowing how good is a classifier based on weak labeling at identifying sentences bearing information on public health events in blog posts. Second, we are interested in characterizing the conditions under which the auxiliary source can be considered a sublanguage for the noisier corpus. We characterize noise based on the length of the sentences or the type of entities appearing in them. We

conduct experiments from two perspectives of the bootstrapping process, namely training and testing phases. Experimental setting and results are described in the next section. At the end, we will discuss the approach.

4.1 Experimental Setting

In our experiments, we use data collected from ProMED-mail [12] as the auxiliary domain. ProMED-Mail (referred to as Promed) is a global electronic reporting system, which lists outbreak reports of emerging infectious diseases. It constitutes a terse source of information about epidemic events. Similarly, the World Health Organization also reports disease outbreak news on their webpage (`http://www.who.int/csr/don/en/index.html`). This data is considered as yet another moderated data source, referred to as WHO.

The data of the target domain is provided by the AvianFluDiary (`http://afludiary.blogspot.com/`). All data was collected directly from the websites. Summary statistics for the data are shown in Table 1.

Source	Years	No. of Documents	No. of Sentences
AvianFluDiary	2006-2009	4249	100890
ProMed-Mail	2002-2009	13369	22170
WHO	1996-2009	1531	16213

Table 1. Data Collection for Experiments

The goal of our experiments is to identify information bearing sentences in AvianFluDiary blogs. As can be seen in Table 1, even for a single blog, spanning less than half the number of years for each moderated source, the number of blog sentences is still over three times greater. We therefore seek to evaluate, the effectiveness of a weakly labeled classifier at detecting relevant disease reporting sentences in such a voluminous and more verbose data source.

In order to do so, we train a classifier on data material of our auxiliary domain (Promed) based on the structural SVM implementation of SVM-TK v1.2 [13]. The features used in the classifier are the tree structure of the parts of speech for each sentence. To create these features, the text was normalized to remove extraneous symbols. Sentence splitting and parse trees were created by the Stanford Parser. Named entities were recognized by applying OpenCalais (`http://www.opencalais.com/`).

As initial training documents for training the classifier, we created a weakly labeled set of examples for the auxilary domain using the automatic labeling approach introduced in section 3.1. In more detail, the top-1 sentences in the sentence database were labeled as positive cases and the bottom-1 as negative examples for our classification task. Observing that not all positive and negative sentences were of equal quality, we prefiltered all top-1 sentences based on

the sentence length. We used default values where the minimum and maximum sentence lengths were set to 20 and 200, respectively.

To test the classifier, a total of 5,029 (729 positive and 4,599 negative) AvianFluDiary sentences were hand labeled. The sentences were labeled with respect to the task of identifying *victim reporting* sentences. The definition we used for victim reporting is based on the MedISys Disease Incidents template[1]. The template reports *disease, time, location, status, cases, description* and *url* to a natural language text.

Therefore, we label a sentence as victim reporting sentence if it contains: D = {disease} in union with I = {time, location, status, cases}. *Status* reports the condition of a victim (e.g., *hospitalized, dead*), and *cases* refers to the number and type of victims (e.g., *bird, child*, etc). Further, we labeled all sentences needed to report a single event as a positive case with the value 1; all other sentences are labeled with 0.

4.2 Experimental Results

The objective of the experiments were two-fold: Determining the quality of the introduced classification approach (Part I), and characterizing conditions for sublanguages (Part II).

Part I: Bootstrapping with Automatically Labeled Data For the first part of the experiments we were interested in determining the quality of a classifier at identifying information bearing sentences when it is trained on weak labeled training material of an auxiliary domain. We tested the classifier on the AvianFlu-Diary data. The bootstrap learner takes into account different scenarios of the bootstrap process.

- **Scenario 1:** Default settings based on the values used by the authors for a similar task (auxiliary domain: Promed),
- **Scenario 2:** Applied on bottom-1 sentences only that are additionally filtered (auxiliary domain: Promed),
- **Scenario 3:** Scenario 1 with WHO as auxiliary data,
- **Scenario 4:** Sentences are filtered based on presence of named entities (auxiliary domain: Promed)

In Scenario 1, the pool size was set to 15, and 50 sentences per iteration was used. A stopping condition was reached when 2,000 items in the weakly labels set were labeled (model size). A classified sentence was selected as confident if its confidence value exceeded 70% percent of the maximum confidence value relative to the given iteration. The initial pool size of 50 positive and 50 negatives sentences was used.

In Scenario 2, similar parameters are used, except that the model size is reduced to 1700. In Scenario 3, WHO data is used as an auxiliary source of data,

[1] `http://medusa.jrc.it/medisys/helsinkiedition/all/home.html`

to see if applying another weakly labeled sentence database yield to different results. Finally, in Scenario 4, we filter the sentences based on the presence of named entities which contain both a medical condition and location. For the different scenarios, the precision, recall and accuracy values of the learner on the AvianFlu-Diary data is examined (see Table 2).

Scenario	Precision	Recall	Accuracy
1	.77	.45	.57
2	.71	.66	.69
3	.75	.22	.34
4	.80	.40	.53

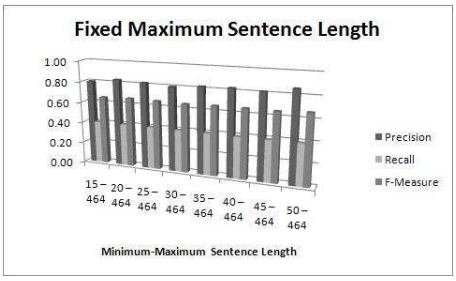
Table 2. Bootstrap Results per Scenario

It can be seen that for the different scenarios differing accuracy values are achieved. The best accuracy of .69 is determined for scenario 2, while the worst accuracy of .34 is achieved for scenario 3. Precision values lie between .71 and .8 for all four scenarios. The recall is significantly lower with values between .22 and .66. We discuss these results in Section 4.3.

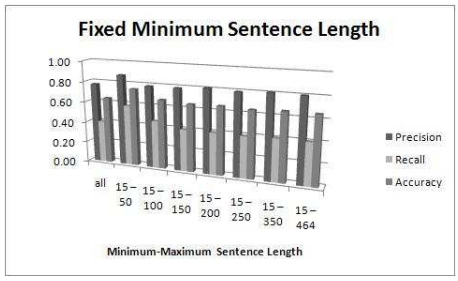
Part II: Analysing Sublanguage Conditions In the second part of experiments, we are interested in characterizing the conditions under which the language of an auxiliary data source can be considered a sublanguage for the noisier corpora. We account for noise by filtering the length of the sentences in the target data and filter based on the type of entities appearing in the sentences. The training sentence lengths were filtered with a minimum sentence length of 30 and a maximum sentence length of 200. The results are shown in Figures 3. When varying the minimum sentence length, the best precision of more than 80% is achieved for sentences with a minimum length of 50. For shorter sentences, the precision drops below 80%. Further, when varying the maximum length of sentences, the best results are achieved for a maximum sentence length of fifty words. In the next section, these results will be interpreted and discussed.

4.3 Discussion

In part I of the experiments (see section 4.2), an important observation is that given a threshold value of .7 for precision, ProMed-Mail can in fact be used in an automatic way to build a bootstrap classifier for blogs. This implies that by using the top-1 sentences as positive cases we are capable of identifying relevant health related sentences in blog postings. Given the fact to no human effort was incurred for labeling a training set, a fairly good classifier can be built based on weak labeling for identifying sentences bearing information on public health events.



(a) Fixed Maximum Sentence Length



(b) Fixed Minimum Sentence Length

Fig. 3. Part II: Analysis of Sublanguage Conditions

We also notice that except for Scenario 2, in which additional filtering was applied to the bottom-1 sentences, the recall tends to be quite low. This would suggest that although top-1 positive cases perform well, the bottom-1 negative examples used in training are not representative enough to distinguish the negative examples present in the blogs. In light of this, we propose that further experiments are needed to better characterize the conditions under which the auxiliary source can be considered a sublanguage for the noisier corpora for identifying negative cases. In part II of the experiments 4.2, we seek notice that although for most of the different sentence lengths, the same results in precision and recall are achieved, in a single range for both the fixed upper and fixed lower, we achieve a noticeable peak precision above 80%. This suggests that such an approach is sensitive to the length of sentences in the test data.

In summary, the introduced approach for sentence classification has been proven to provide acceptable results. In contrast to existing approaches, the main benefits of the method presented here are: (a) Avoidance of manual labeling of training material, and (b) Reduction of bias produced by automatic labeling through bootstrapping for learning the classifier. The presented approach is new for several reasons: Bootstrapping has not yet been applied for learning a classifier in a cross-corpora setting for labeling sentences. The problem of victim-reporting sentences classification was only considered using pattern-based approaches by now. We introduce a weakly-supervised approach to address this problem. Further, the automatic labeling of examples and its combination with bootstrapping to reduce incertitude has not been reported and analysed before.

5 Conclusion

In this work, an approach is described to identify disease- and victim-reporting sentences from blogs to support epidemic intelligence. Challenges given by the characteristics of blogs (mainly noise and data abundance) are addressed using automatic labeling of data collected from moderated sources to learn a classifier for identifying relevant sentences present in blogs. A bootstrap process is applied

to learn a classifier and filter more noisy and irrelevant sentences. The results show that the approach taken here is quite effective at sentence level filtering in blogs. Without manual effort, we are able to achieve a precision as high as .80 and a recall .66. In the future, we will perform robust experiments, particularly using more blogs. We also intend to experiment with increasing values for top-N and compare this to bootstrapping on the blog set alone and using a hybrid approach.

References

1. C. Paquet, D. Coulombier, R.K., Ciotti, M.: Epidemic intelligence: A new framework for strenghtening disease surveillance in europe. Euro Surveill. **11(12)** (2006)
2. Moens, M.F.: Information extraction from blogs. In Jansen, B.J., Spink, A., Taksa, I., eds.: Handbook of Research on Web Log Analysis, IGI Global (2009) 469–487
3. Yangarber, R.: Verification of facts across document boundaries. In Proceedings International Workshop on Intelligent Information Access (2006)
4. Freifeld, C.F., Mandl, K.D., Reis, B.Y., Brownstein, J.S.: Healthmap: Global infectious disease monitoring through automated classification and visualization of internet media reports. In Proceedings International Workshop on Intelligent Information Access (2006)
5. Collier, N., et al.: Biocaster: detecting public health rumors with a web-based text mining system. Bioinformatics, Oxford University Press (2008)
6. Mykhalovskiy, E., Weir, L.: The global public health intelligence network and early warning outbreak detection: a canadian contribution to global public health. Can J Public Health **97(1)** (2006) 42–44
7. Wang, B., Spencer, B., Ling, C., Zhang, H.: Semi-supervised self-training for sentence subjectivity classification. Canadian AI 2008, LNAI 5032 (2008) 344–55
8. Didaci, L., Roli, F.: Using co-training and self-training in semi-supervised multiple classifier systems. In: D.-Y. Yeung et al. (Eds.): SSPR&SPR 2006, LNCS 4109 (2006) 522–530
9. Chen, Z., Ji, H.: Can one language bootstrap the other: a case study on event extraction. In: SemiSupLearn '09: Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing, Morristown, NJ, USA, Association for Computational Linguistics (2009) 66–74
10. Zhen, Y., Li, C.: Cross-domain knowledge transfer using semi-supervised classification. In: AI '08: Proceedings of the 21st Australian Joint Conference on Artificial Intelligence, Berlin, Heidelberg, Springer-Verlag (2008) 362–371
11. Yangarber, R., et al.: Combining information about epidemic threats from multiple sources. In: RANLP-2007, Borovets, Bulgaria (2007)
12. Madoff, L.C.: Promed-mail: An early warning system for emerging disease. Clinical Infectious Diseases **2**(39) (July 2004) 227–232
13. Moschitti, A.: A study on convolution kernels for shallow semantic parsing. In: ACL '04, Morristown, NJ, USA, Association for Computational Linguistics (2004) 335

Linking Specialized Online Medical Discussions to Online Medical Literature

Sam Stewart[1⋆], Syed Sibte Raza Abidi[1], Allen Finley[2]

[1] NICHE Research Group, Faculty of Computer Science, Dalhousie University, Halifax, Canada
[2] IWK Health Centre/Dalhousie University, Halifax, Canada

Abstract. The medical web comprises both medical communities engaged in discussions about specialized topics and a vast array of medical articles available through web-based databases. In this paper we present a knowledge linkage strategy that links online specialized medical discussions with corresponding online medical articles. The idea is to link the experiential knowledge generated in online medical discussions by a virtual community of specialized medical practitioners with the explicit knowledge available in online medical literature archives. We have developed a specialized medical literature search algorithm, based on the principles of the Extended Boolean Information Retrieval algorithm [6], to retrieve a ranked list of medical articles associated with the specialized medical discussion. The medical literature search algorithm is part of our knowledge linkage strategy that involves the generation of topic-specific discussion threads from online discussions, formulation of highly specialized search queries based on a specialized discussion thread and retrieval of published medical articles from PubMed that are closely related to the online discussion. We have applied our knowledge linkage strategy to the specialized medical topic of Pediatric Pain Management, and have achieved an improvement in the positive return rate (recall) from 55% to 70% in terms of linking online medical discussions to the correct medical articles.

1 Introduction

Web 2.0 technologies are embraced by medical practitioners for collaborative case solving, professional communications, knowledge sharing, medical education, patient interactions and so on. From a knowledge and experience sharing perspective, online discussion forums and mailing lists provide a viable medium for medical professionals to virtually engage in discussions around specialized medical topics. The ensuing discussions, which constitute a thread of emails or postings by medical professionals from different parts of the world with varying degrees of expertise and experience, entail practical know-how in terms of what worked and what did not work, recommendations and solutions to unusual cases

⋆ This work is carried out with the aid of a grant from the International Development Research Centre, Ottawa, Canada

and references to domain experts or published evidence. Such online discussions are a vital resource for experiential medical knowledge emanating from a community of medical practitioners. Notwithstanding the utility of online discussions on specialized medical topics, medical practitioners like to correlate the recommendations with published medical literature for use in clinical decision-making. The process of searching for information within specialized domains, however, is a key challenge within the medical community. Studies have shown that the lack of clinical knowledge about specialized subjects, such as pediatric pain, have lead to incorrect interventions [2]. These problems tend to be exacerbated by the fact that specialized practitioners do not often have the time to meet and share information face-to-face, forcing them to rely on their own search strategies to retrieve information from published resources.

Linking specialized online medical discussions to online medical literature poses an information retrieval challenge because the specialized discussions are context-sensitive, spanning multiple emails/postings and encapsulate concepts from multiple sources. A medical practitioner seeking medical articles corresponding to the online discussion is therefore required to formulate a *focused* search query that captures the discussion's context, uses the prevalent terms and is posted to the right online medical literature archive. It may be noted that the process of finding research articles related to a specialized topic amongst the nineteen million different articles available on the online database of Pubmed is a challenging task, and can be even more challenging for specialized fields for which there is less published literature. This paper presents a medical literature retrieval strategy that automatically comprehends a specialized online discussion to formulate a search query that retrieves relevant medical articles from a web of medical literature archives (in particular the online databases of Pubmed). In this way, we establish *knowledge linkages* between the experiential knowledge encapsulated within online medical discussions with explicit knowledge stored in online medical literature archives.

In this paper we present our knowledge linkage strategy that involves a sequence of steps, starting from forming topic-specific discussion threads to formulating highly specialized search queries based on a specialized discussion thread to retrieving a set of published articles from PubMed that are closely related to the online discussion (see figure 1). We have developed a specialized medical literature search algorithm, based on the principles of the Extended Boolean Information Retrieval algorithm [6], that incorporates both weighted and unweighted query terms (keywords derived from the selected medical discussion) to retrieve a ranked list of medical articles associated with the specialized discussion. We use Metamap [1], a program designed by the National Library of Medicine for processing the free-form medical text of the discussions to a set of medical keywords based on the MeSH lexicon. The choice of MeSH terminology is quite natural since the PubMed data is indexed by MeSH keywords. We have applied our knowledge linkage strategy to the specialized medical topic of Pediatric Pain Management that features a Pediatric Pain Mailing List (PPML) with over 700 subscribers. Our results show that the application of our special-

ized medical literature search algorithm has improved the positive return rate (recall) from 55% to 70% which is a significant improvement in terms of linking online medical discussions to the right medical articles.

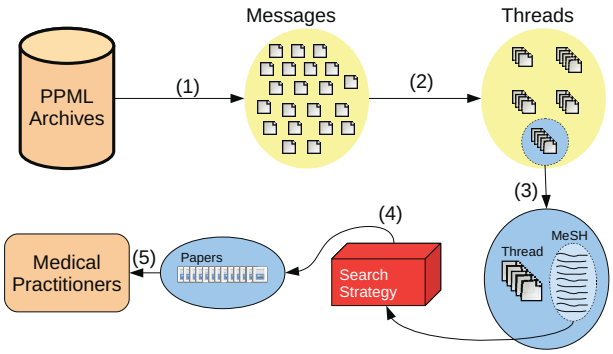


Fig. 1. The knowledge linkage strategy. (1) Messages are extracted and (2) combined them into threads, where they are then (3) linked to formal medical terms. These terms are then (4) used in a novel search strategy to obtain a ranked list of papers, which are then (5) returned to the practitioners.

1.1 Pediatric Pain

Pediatric pain management is an example of a specialized medical domain that can benefit from knowledge linkage. Pediatric pain is a complex subject that is dispersed across multiple departments within a hospital. It is difficult to manage, as children lack the ability to properly express their pain [2], which can lead to incorrect interventions. To compound the problem, healthcare practitioners do not receive proper training in the management of pediatric pain [3], and the multidisciplinary nature of the subject makes it difficult for pediatric pain practitioners to meet and discuss their issues face-to-face.

The PPML is an example of a web 2.0 tool that has provided an electronic link between clinicians working in different departments and hospitals around the world. The PPML has over 700 subscribers and over 13,000 messages, all archived, making it an excellent candidate for knowledge linkage. The conversations on the mailing list will be processed using the program Metamap, which will provide a list of pertinent medical keywords extracted from the MeSH lexicon.

1.2 Metamap

Metamap is a Natural Language Processing (NLP) tool designed to parse free-form medical text and connect it to formal medical terms from selected medical

lexicons. For this project the lexicon being used is the MeSH vocabulary, but Metamap has the ability to map to any lexicon in the Universal Medical Language System (UMLS), such as SNOMEDCT or ICD9. Metamap has been used in several other projects to link free form medical texts to formal medical terms [4, 5]. For more details on its use see Aronson's introductory work [1].

2 Methods

The objective of the search strategy is to *passively* link the conversations on the mailing list to pertinent published literature. This means that the MeSH terms produced by the mapping process and their scores must be leveraged by the search strategy to produce a ranked list of papers associated with that conversation. Other projects that have looked to make information retrieval in the medical domain easier have looked at ways to improve clinicians *active* search strategies, through better search algorithms and interfaces [8]. This project takes a different approach, choosing to perform the search *automatically* without requiring clinician input. The resulting set of papers will be provided without requiring any input from the user, vastly increasing the speed of the knowledge linkage process. If the resulting set of papers is not optimal then the set of ranked MeSH terms returned can be used to inform a manual search.

2.1 Search Strategy

The search strategy is based on the Extend Boolean Information Retrieval (eBIR) algorithm developed by Salton et al [6]. The algorithm builds on the traditional boolean information retrieval approach by including both query and document weights for each of the keywords, and then using a p-norm calculation to assign a search score. This project will modify the eBIR algorithm to better fit automatic searching within specialized domains.

2.2 eBIR and p-norms

Boolean information retrieval is the simplest form of information retrieval, in which query terms are joined by AND and OR operators, and any papers matching the query are returned. There a several problems with the boolean information retrieval model. First, it is often difficult to manage the size of the returned set of papers; complex searches can easily return no papers, yet removing a search term can result in a set of several thousand papers. Second, there is no ranking of the papers returned. Third, there is no way to assign importance to specific keywords. Finally, there is a problem with the structure of the searches; if ten query terms are join by AND operators, then papers that match nine of the terms but not the tenth are not returned. In the context of this project the boolean search strategy is particularly ineffective, as it does not make use of the Metamap scores at hand, and there are far too many query terms associated with a conversation to retrieve a manageable set of papers.

To remedy this problem Salton et al. developed a system that incorporates term weights to aid in the search process. The eBIR algorithm allows weighting of both the paper keywords and the query terms. For this project there are no weights for the document keywords (which are assigned by the authors manually via Pubmed), but the Metamap scores can be used as query weights, with higher weights indicating more confidence in the search term. Though the eBIR algorithm suggests that weights should be in the range of [0,1], there is no reason mathematically that they cannot be in the range of [0,∞], and thus no transformation of the Metamap scores is required.

The eBIR algorithm uses the idea of p-norms to measure the score of a set of OR or AND terms. Let the set of query terms be represented $A = \{(A_1, a_1), \ldots, (A_n, a_n)\}$, where A_i is the i^{th} query term, and a_i is the associated score. Let a document D be represented by the set $D = \{d_{A_1}, d_{A_2}, \ldots, d_{A_n}\}$ where d_{A_i} is the weight associated with keyword term i in that specific document. Since this project does not allow for weighted document keywords $d_{A_i} = 0$ or 1. Let the query $Q_{OR(p)} = \{(A_1, a_1)\ \text{OR}^{\,p} \ldots\ \text{OR}^{\,p} (A_n, a_n)\}$ by the set of query terms linked by OR, and let the query $Q_{AND(p)} = \{(A_1, a_1)\ \text{AND}^{\,p} \ldots\ \text{AND}^{\,p} (A_n, a_n)\}$ by the set of query terms linked by AND. The p-norm scores for each of the searches is given in equations (1) and (2).

$$sim(D, Q_{OR(p)}) = \left[\frac{a_1^p d_{A_1}^p + a_2^p d_{A_2}^p + \ldots + a_n^p d_{A_n}^p}{a_1^p + a_2^p + \ldots + a_n^p}\right]^{1/p} \tag{1}$$

$$sim(D, Q_{AND(p)}) = 1 - \left[\frac{a_1^p (1 - d_{A_1})^p + \ldots + a_n^p (1 - d_{A_n})^p}{a_1^p + \ldots + a_n^p}\right]^{1/p} \tag{2}$$

The selection of p effects the relative strengths of the returned scores. Selecting $p = \infty$ results in a standard boolean information retrieval model, while selecting $p = 1$ results in a vector-space model [7], in which the ANDs and ORs are ignored and the papers are ranked by the sum of the query terms that appear in each paper.

For this project the simplest form of an eBIR algorithm would be to link all the terms returned by Metamap using an OR operator. Let the set $M = \{(M_1, m_1), (M_2, m_2), \ldots (M_n, m_n)\}$ be the MeSH terms and their scores for a particular conversation. Then the query would be given in equation (3), and the score calculation for paper D would be given by equation (4).

$$Q_{OR} = [M_1\ \text{OR}\ M_2\ \text{OR}\ M_3 \ldots M_n] \tag{3}$$

$$sim(D, Q_{OR(p)}) = \left[\frac{m_1^p d_{M_1}^p + m_2^p d_{M_2}^p + \ldots + m_n^p d_{M_n}^p}{m_1^p + m_2^p + \ldots + m_p^p}\right]^{1/p} \tag{4}$$

Note that the selection of p is key to the function of the p-norm calculation and subsequently the eBIR algorithm. Setting $p = 1$ makes sense theoretically, as the principle behind the $OR(p)$ operator is to return the papers that match the most number of terms in the query set, so equation (4) could be reduced to

$sim(D, Q_{OR}) = \sum m_i d_i$, where d_i is an indicator of whether term i is a keyword for the paper.

The problem with the eBIR algorithm is that it is not well suited for specialized domains. The Metamap program extracts keywords that represent the conversation within the mailing list, but keywords such as *Pediatrics* and *Pain* are implicitly representative of all conversations on the list, whether or not they are particularly suited to the conversation. This problem needs to be addressed, to make sure that the search strategy is focusing on the correct body of literature.

2.3 Modified Information Retrieval Algorithm

To solve the problem of specialized domains it was decided that a *specialized filter* would be added, adding an AND operator to the query. The objective of the specialized filter is to focus the search on papers relevant to the specialized subject. One has to be careful, however, to not over-restrict the search by filtering out useful papers. To this end an age-group filter is added, to ensure that all papers returned are relevant to the pediatric population. The new query would modify equation 4 by adding *Infant*, *Child* and *Adolescent* to the set of MeSH terms, as demonstrated in equation (5).

$$Q = [Infant \text{ OR } Child \text{ OR } Adolescent] \text{ AND } [M_1 \text{ OR } M_2 \text{ OR } M_3 \ldots M_n] \quad (5)$$

If the eBIR algorithm were used then the next step would be to apply query weights to the terms in the specialized filter and then find a suitable value for p. This project decided instead to modify the eBIR algorithm slightly, by combining the idea of strict boolean searching with a weighted query.

The final search algorithm leverages the eBIR idea of weighting queries, but adds a strict filter that reduces the search field to only those papers that match the age filter. This filter has the effect of focusing the search strictly on papers that focus on the pediatric population. The score for paper D is therefore calculated using the equation 6. The equation uses the same calculation as the eBIR algorithm, but requires the presence of one of the age group keywords. Let d_I, d_C and d_A be the indicators of whether the paper contains the MeSH terms *Infant, Child* or *Adolescent* respectively.

$$sim(D, Q) = [1-(1-d_I)(1-d_C)(1-d_A)](m_1 d_{M_1} + m_2 d_{M_2} + \ldots + m_n d_{M_n}) \quad (6)$$

3 Results

This is an example of a single conversation from the PPML.

> **Question:** *We are looking at ways to decrease the pain of ocular flushing necessary when a child gets sand or spray in their eyes. I am really having a hard time finding any literature on what is the most comfortable solution to use*

(NS?RL)and what freezing drops to use(Prilocaine?) If anyone has a procedure, or protocol, or literature to share or can let know what you are using, I would really appreciate it. Right now we are using nothing.
Response: *From personal experience, Proparacaine (Alcaine(r) in the U.S.) anesthetic eye drops are almost painless on instillation; they do not provide as deep anesthesia as tetracaine but burn much less and usually provide sufficient conjunctival and corneal anesthesia. ...*

A sample of the MeSH terms associated with this conversation are available in table 1, and seem to be a reasonable representation of the conversation. The full set of MeSH terms was used in the search strategy to retrieve the set of papers, the top two of which were as follows:

- Boscia F et al. *Combined topical anesthesia and sedation for open-globe injuries in selected patients.* Ophthalmology:2003,110(8).
- Snir M, et al. *Efficacy of diclofenac versus dexamethasone for treatment after strabismus surgery.* Ophthalmology:2000,107(10).

These papers seem to be pertinent to the subject being discussed. This example demonstrates the effectiveness of the system, and its ability to provide published literature to supplement the information being shared online.

MESH	**Score**
Lenses	2583
Pain	2434
Anesthesia	1722
Anesthesiology	1722
Eye	1000

Table 1. A sample of the mappings corresponding to the sample conversation

3.1 Evaluation

There is a challenge in evaluating a search strategy of this type. The strategy is specific to unstructured, medical conversations, and it is therefore not possible to apply the search strategy to traditional annotated information retrieval databases. Without an annotated database it is difficult to calculate precision and recall, the traditional measures of information retrieval systems. An alternative strategy for evaluation is therefore required.

The search strategy was tested on a sample of conversations from the PPML between 2007 and 2008. For each conversation Metamap was used to map the conversation to a set of MeSH terms. The MeSH terms were then fed to both search strategies (the eBIR algorithm and the modified algorithm), and the top 15 papers returned by each search were linked to the appropriate thread. The

threads were evaluated to see if the set of papers returned was appropriate. A set of papers was deemed appropriate if at least one of the returned papers was relevant to the subject being discussed.

The results of the study were promising. For the eBIR algorithm 55% of the papers returned were deemed relevant to the thread. This percentage jumped up to 70% for the improved algorithm, a significant increase over the first attempt ($p = 0.0025$). The improvement is due to the filter, which restricted the search area to those papers relevant to the pediatric population.

4 Conclusion

The purpose of *knowledge linkage* is to provide clinicians with quick access to evidence-based knowledge to supplement the tacit knowledge they share via web 2.0 communications. Because the information retrieval process is done passively, without clinician input, a robust algorithm is required that can consistently return pertinent medical knowledge. This paper has presented an algorithm that incorporates query weights to automatically produce a search query that is appropriate for specialized knowledge domains such as pediatric pain. The algorithm was built on the eBIR algorithm, and has been proven to significantly improve the relevance of the papers returned.

Future research should be directed at a larger study of the two algorithms, along with a comparison to a more sophisticated eBIR implementation. A better evaluation of the search strategy should be completed, including evaluating the precision and the recall of the strategy, and an implemented system should be tested to evaluate the overall usability of the system.

References

1. Aronson, A.R.: Effective mapping of biomedical text to the umls metathesaurus: The metamap program. Proceedings of the AMIA Symposium (2001)
2. Atherton, T.: Children's experiences of pain in an accident and emergency department. Accident and Emergency Nursing 10, 79–82 (1991)
3. Caty, S., Tourigny, J., Koren, I.: Assessment and management of children's pain in community hospitals. Journal of Advanced Nursing 22(4), 638–645 (1995)
4. Chapman, W.W., Fiszman, M., Dowling, J.N., Chapman, B.E., Rindflesch, T.C.: Identifying respiratory findings in emergency department reports for biosurveillance using metamap. MEDINFO (2004)
5. Chase, H.S., Kaufman, D.R., Johnson, S.B., Mendonca, E.A.: Voice capture of medical residents' clinical information needs during an inpatient rotation. Journal of the American Medical Informatics Association 16, 387–394 (2009)
6. Salton, G., Fox, E., Wu, H.: Extended boolean information retrieval. Commun ACM 26(11), 1022–1036 (1983)
7. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Commun ACM 18, 613–620 (1975)
8. Trieschnigg, D., Pezik, P., Lee, V., de Jong, F., Kraaij, W., Rebholz-Schuhmann, D.: Mesh up: effective mesh text classification for improved document retrieval. Bioinformatics 25, 1412–1418 (2009)

The Importance of RSS in the Exchange of Medical Information

Frankie Dolan and Nancy Shepherd

[1] MedWorm.com
frankie.dolan@medworm.com
[2] Shepherd Research LLC.
nancy@shepherdresearch.com

Abstract. This paper investigates the role of RSS in providing a solution to the problem of medical information overload, speeding up the dissemination of information and improving communications between all those with an interest in health. It compares the exchange and use of medical information on the Internet before and after the use of RSS and also shares a vision for the future, using MedWorm, a medical search engine and RSS newsfeed provider, as an example. The conclusion highlights how RSS has opened a new dimension of information exchange which has the potential to enable giant steps forward in the field of medicine. To realise its full potential, both publishers and users of medical information need to recognise the importance of RSS, ensure thoughtful implementation of RSS feeds to announce publication, and provide for education regarding its everyday use.

1 Introduction

The Internet has enabled access to a wealth of in depth research and medically related information not previously available, but it has also given rise to a new set of problems for todays physician. Medical practitioners are now inundated with information[1], short of time [2] and yet obliged to keep up to date at all times with the very latest developments. Patients are researching their own conditions and often expect their doctors to have expert and recent knowledge on a vast range of topics.

This paper briefly describes RSS (really simple syndication) [3] and investigates the way in which RSS is starting to provide a solution to the problem of medical information overload, speeding up the dissemination of information across the Internet and improving communications between all those with an interest in health.

Since RSS has been around for several years now, its benefits and situations in which it is appropriate are already known. This paper considers its uptake and re-emphasizes its importance, particularly in the field of medicine where the delivery of timely and highly relevant information is essential. It compares the exchange and use of medical information on the Internet before and after the use of RSS and also shares a vision for the future.

This paper uses MedWorm as an example. MedWorm is a medical search engine and RSS newsfeed provider based on data collected from over 7,000 medical RSS feeds [4].

2 Background

2.1 RSS

RSS is a group of simple data formats that are used to announce new data on the Internet. The first version of RSS was created in 1999 although it wasnt until 2005 to 2006 that it started to gain widespread use [5]. The most commonly used format of RSS today is the latest version, RSS 2.0 [6] which is popular due to its simplicity. An alternative to the RSS format is Atom [7]. Applications used to read RSS feeds are commonly known as RSS readers or aggregators, and RSS is also often referred to as aggregation. Many RSS aggregators, including MedWorm, can parse data from any of the different RSS formats. An overview of the advantages and disadvantages of RSS is presented in Table 1.

2.2 Uptake

RSS implementation has been growing steadily within the field of medicine since its introduction in 2006. A few examples of its use are described below.

It is now common practice for all leading medical publishers to include RSS feeds for each of their medical journals. RSS feeds can also be pulled out from any search run via PubMed [8]. The Ebling Library [9] has compiled a feed directory of over 3,000 medical journals categorised by medical specialty. Making extensive use of RSS are medical websites such as Medscape [10], health consumer websites such as WebMD [11], medical news sites such as Medical News Today [12], all leading newspapers and news services such as Reuters [13]. RSS is also being used effectively in several instances to announce industry alerts as seen by the FDA [14] and recently by the WHO [15] and Flu.gov [16] to keep people up to date with the latest swine flu reports. The CDC [17] uses RSS to announce its morbidity and mortality weekly reports and the NHS Library [18] provides RSS feeds for all of its categories.

Medical universities are now often including RSS feeds for their latest news updates and some, such as Duke University [19], are even offering specialised RSS training on request. RSS is frequently covered as a topic in medical research papers and presentations. The University of Helsinki has developed FeedNavigator [20] which is an online medical aggregator. RSS feeds are also now frequently used on medical websites to display updates from other sources and are published automatically on nearly all medical blogs.

2.3 Room for Improvement

The uptake of RSS within the field of medicine has been widespread within a relatively short space of time. However, the majority of medical professionals,

unless they are keen adopters of Internet and the latest information technologies, have yet to learn about or appreciate the benefits of RSS, and have not yet widely adopted the use of personal RSS readers or personalized RSS aggregation services. In addition, there are still many medical websites not posting RSS feeds for key data updates, in particular for other data types that are not classed as medical journals or news.

3 Past and Present

3.1 Data Collection

In the past, without the use of RSS, an online information distribution service would receive data from different sources using various transfer methods such as email, disc, or online download. Incoming data came in various formats, using different types of identifying tags, if at all. Internal programming teams were often required for the writing of data conversion programs to convert the inconsistent data received into a standard format that had been adopted by the data distribution company. The whole process was labour intensive with different publishers requiring customised data conversion programs. Usually the full data in its entirety was copied across to a central database held by the data distribution company.

Today, using RSS, publishers simply update an RSS text file online when new information is released. A data distribution service that uses RSS, such as MedWorm, can automatically check the RSS files regularly for updates. Data is tagged with standard RSS formatting which means that customized data conversion programs are no longer needed, instead only one RSS data parser is required. Providing concise summaries with hyperlinks to the full text of articles means it is no longer necessary to submit data to a central database. The full text of the data can now remain at the point of origin with the RSS feed items acting as pointers.

3.2 Search Engines

The comparison of a search engine based on data collected from web crawling to that of one based on data gathered through syndication may not at first glance seem a valid exercise, since a search engine such as Google indexes the content of web pages themselves, whereas an engine relying on RSS feeds focus only on updates. However, from the end user's perspective, the end result and purpose of the search engine looks similar, and in medicine both are used for the same purpose, to get information to answer questions. End users frequently ask for clarification on why they should use any other search engine to that with which they are familiar. It is therefore important that the informaticist clearly understands the benefits of the RSS populated database if they are to encourage users to make use of such a resource.

Without the use of RSS, a search engine such as Google uses a Web crawler to find information to populate its database by following hyperlinks from one page

to another. All data found is cached and indexed. The possibility of duplicate date is large since webpages can have many different formats and yet contain the same content, making it difficult for duplicate data to be identified within the database. Descriptive data is collected from metatags, which may or may not have been included, and which are prone to manipulation for search engine optimization techniques. It is sometimes difficult to determine any clear title and frequently impossible to identify any summary text. Most limiting of all is the common lack of a distinct publication date. In many cases the most accurate date that the Web crawler can provide will be the date on which the article is discovered. Web crawlers waste a lot of processing power by revisiting pages looking for new content and it can take a long time for new pages to be discovered and indexed. In addition, the amount of data to be included can have an impact on performance and slow down the time it takes to get relevant and up to date content indexed for the audience to find.

Now with RSS, a search engine such as MedWorm uses RSS feeds to populate its database with benefits over that of a Web crawler. All items include a clear title and published date, and often there is also provided a concise summary of the publication. Data is less prone to manipulation by keyword cramming, since if the content does not make sense, the user will unsubscribe. Using RSS, when unformatted text and article summaries instead of full text are indexed, processing is quicker [21]. This enables frequent re-indexing to ensure that queries always return the most up to date information. In addition, processing power is not wasted from the revisiting of pages to look for updates, since it is understood that newly published data will be announced through a new RSS item that will be automatically indexed for inclusion.

For the user of medical information, the greatest advantage that an RSS based search engine has to offer over a standard Web crawled search engine is the speed at which data is indexed and the ability to return search results by date order. Up to date information can often be more important than link popularity in the context of medically related information. Medical professionals and scientific researchers need to place information in the context of a publication date and in relation to other information known previously.

3.3 Subscription to Data

The traditional method of subscribing to data is via email update. Email communications suffer from numerous problems, including spam [22] and misdirection. Providing updates via email subscription is known as push technology, since the publisher sends out the emails and could keep sending them whether the recipient wanted to receive them or not.

RSS subscriptions are known as pull technology since the user chooses which RSS files they wish to receive and how frequently the information will be pulled. If a recipient chooses to unsubscribe, they simply remove the link to the RSS file that they no longer wish to include in their RSS aggregator, which eliminates the problem of spam.

3.4 Sharing Data Between Sites

To share data from one website to another it used to be necessary to either use a customised plug-in or to learn how to program an API that had been designed specifically for that site. Such an approach can be difficult to implement requiring programming skills, or can be limiting.

With RSS it is relatively simple to install one of many RSS widgets that can read RSS files and output the contents in many different formats. Data can be set to cache and refresh as wished to ensure that data is always up to date but bandwidth is saved.

Registered MedWorm users may create customized RSS feeds to place on their own websites. For this, complex queries may be combined and then optionally manually filtered further if need be. Descriptions and homepage links for the RSS feeds can also be set and there is the option to track which feed items have been read from the customized feeds.

4 Looking to the Future

The benefits of RSS as discussed above are well known within the IT community and the prevalence of RSS, particularly within the field of medicine, is growing.

Here we list some areas that have been identified through experience with MedWorm as having potential for future development.

4.1 Clinical Trials

Although clinical trial data may be found online, in particular in the central repository ClinicalTrials.gov [23], it can be difficult for physicians and their patients to keep up to date with clinical trials that are relevant. It is now possible to create an RSS feed for clinical trials for any category via ClinicalTrials.gov and MedWorm is considering how it might integrate such data. If all sites that published data relevant to clinical trials were to include RSS feeds for such data, this would help improve communication.

Using RSS to announce new clinical trials does not solve the problem of complexities of data structure or provide reporting functionality to cater for different terminology or search by eligibility. However, it does provide a very simple way of keeping physicians constantly and broadly aware of new opportunities within their field as they arise. Having clinical trial feeds indexed by MedWorm would enable trial updates to easily reach their target audience, since medical professionals and patients already use MedWorm to get daily updates on topics of their interest, either via visiting the site or through email or RSS subscription.

4.2 Funding Opportunities

Currently many worthy organisations, medical scientists and physicians miss out on vital grants and other funding possibilities, often finding out about them

only after a closing date has passed. Email subscription is commonly used but as described previously this method has its weaknesses. MedWorm has been working closely with the Hope Center for Neurological Disorders at Washington University to ensure key funding RSS feeds are included within MedWorm. Utilizing RSS and MedWorm to announce funding updates would ensure that these opportunities reach their target audience in a timely manner.

4.3 Discussions & Commentaries

Medical discussion forums on the Internet are valuable in their facilitation of building new relationships and sharing knowledge. However, forums are often prone to spam and manipulation. Important knowledge can be buried with time and much repetition can take place with the same questions being asked and answered in multiple places. Forums sometimes lack focus and webmasters are often faced with the problem of ensuring enough traffic to keep on stimulating forums with new discussions over time.

Items from an RSS feed could now be used as a basis for discussion, bringing people together to share ideas on the most recent medical research and news. Any query run on MedWorm to pull out an RSS feed of publication items on a particular topic also returns an RSS feed for the discussion items connected to the publication records returned.

In addition, MedWorm registered users may now use the discussion functionality to pull out RSS feeds on just the items on which they have commented. This functionality can be used for medical professionals to start to build up their own running commentaries on topics of their choosing, which could be displayed on their own websites.

4.4 Social Networking, Digital Journal Clubs and Focus Groups

MedWorm now has the potential to link up users reading the same items to other users reading the same data, should users opt in to share their reading habits. RSS use increases readership to levels where such connections start to become meaningful. The opening of new communication lines leads to new ways of individuals working together. MedWorm already provides a platform for discussion on current articles. RSS use has brought about the concept of running a digital journal club, through the selection of articles to review from the RSS database and then the creation of new RSS feeds for reviewed articles.

Similar to a digital journal club, focus groups could be formed to concentrate on particular illnesses. MedWorm provides a novel meeting ground for different user types, from medical scientist through to patient, to discuss common areas of interest but from differing perspectives.

4.5 Trend Identification

Due to the increased exchange of data and readership that RSS provides, MedWorm is starting to identify trends within the field of medicine, and to provide

RSS feeds for these trends. The value of such statistics comes from the combination of the amount and variety of data being parsed, the MedWorm taxonomy and the wide spread use of MedWorm provided data: on the MedWorm site itself, in the RSS feeds that individuals subscribe to via their own RSS readers, and also on other websites that utilize the MedWorm feeds.

Currently reports include the most frequently read items, by each category as well as overall. All reports are provided with their own RSS feed. In the future it will also be possible to obtain feeds for the most frequently arising topics within the data itself, as well as the ability to see which topics have seen increased activity within a specified time.

4.6 An Added Dimension

Just like the advances that came to society with the building of roads, the invention of the telephone and the development of the Internet, RSS provides a new dimension of communication that could see rapid advances being made in medical knowledge. For the first time ever, publishers all across the globe can easily and almost instantaneously route their data to interested audiences throughout all sectors of society. Users will be able to connect to other people reading the same data and exchange thoughts in real time, and those discussions and thoughts will not be lost but will be fed back into the growing body of evidence in which we can all participate.

5 MedWorm Recommendations

RSS promotion within the field of medicine has seen real progress over the past four years. In order to reach the full potential of RSS, there must be a concerted effort from all in the field of medical informatics to further promote the use of RSS in their setting.

5.1 Integration

The first step towards the RSS ideal is to ensure that all new medical information which is meant to be shared broadly gets submitted effectively into the RSS highway.

Here we discuss some of the issues that the developer often faces when looking at how best to integrate RSS with their website:

Data Types When adding any new medical information to the Internet, it should be a priority to ensure that the information is announced by way of a new RSS item posted to an RSS feed, not restricting RSS use to that of medical journals, blogs and news updates. Other types of information that should also post updates to an RSS feed include: all publications that have not been included in the feed of a journal; events such as conferences, exhibitions, and

training courses; clinical trials; online courses; guidelines; press releases; directory items; funding opportunities; podcasts; alerts and recalls; legal outcomes; medical images and videos.

Using a separate RSS feed for each different type of data, as listed above, enables clearer data definition within MedWorm. When several different feeds exist for a site it can be useful to also have one RSS feed that combines all of your RSS releases into one catch all feed, for those subscribers that want to be kept up to date with everything on your site but dont want to have to subscribe to many different feeds separately. However, other than that the one catch all feed, items should not be duplicated across RSS feeds but rather appear in just one RSS feed to help avoid duplicate items appearing in systems that use your feeds for data population.

Intent to Share Including a lengthy RSS terms and conditions legal document can discourage other websites from using your feeds, which could limit your readership. It is better to include content in your feeds that you are happy to be shared anywhere, anyhow, without the need for legal notices. If there is reluctance to share data within your organization, include a description of the data rather than the data itself.

Content Include a descriptive title and date of release as well as a description wherever possible. Ensure the key words relevant to the item are included somewhere within the feed item, without resorting to 'keyword cramming' that would degrade the quality of the data. A concise description of an article is preferable to the whole article itself. It is also better to include a short paragraph that is of use on its own rather than a half sentence 'teaser' that tells the user little.

Do not include feed items that present the user with nothing more than a subscription page when clicked through. If a subscription page is to be used, add some value to the feed by providing some free information when people click through on a feed item, before logon is necessary. Such an approach provides an incentive for readers to continue to subscribe to your feed even when they do not think they want to become a customer, and will likely result in additional sales over time.

Discovery All RSS feeds should be clearly labelled with an orange RSS logo. Any webpage that has an associated RSS feed on it should also include RSS auto-discovery mark-up in its html header [24] There should be an RSS feed on every home page of every medical website. If there are more than one RSS feeds available for a site, there should always be a link from the homepage to a page listing all the feeds available for that site with descriptions, except in the case where the number of RSS feeds for a site is enormous due to dynamically created feeds from different data categories, as seen on MedWorm.

The addition of new RSS feeds for a site should also be announced via an RSS posting on an existing RSS feed on the same site. It is good practice to also

include an OPML file [25] to list all feeds on a site when the number of feeds is numerous. A suggested standard that could be adopted is to include an OPML file for every site with numerous RSS feeds, that includes the file name and location of the file as /rss.opml, similar to the standard of adding a robots.txt file to a site, as suggested by Jeremy Zawodny back in 2003 [26]. At the time the suggestion was not warmly welcomed, but since then it has become common practice to pass lists of RSS feeds from one system to another using the OPML format, and also the number of RSS feeds offered by any one site has typically increased, so this might now be considered as a logical step forward.

Continuation When redesigning a site, give considerable thought to RSS items that have previously been posted to items on the existing site. Ideally any move of data from one link to another should include with it auto redirections for visitors to the previous data links. You may wish to plan ahead for potential future site redesigns by using dynamic urls in your RSS feeds that can be mapped to different physical urls through script.

5.2 Education

The second step towards the RSS ideal is to ensure that end users are trained and encouraged to use RSS to receive data wherever possible.

This should include the provision of RSS training either within the work setting, at medical school, or in continuing education courses. It may be difficult to change the habits of an older generation when it comes to data retrieval and the integration of new technologies to assist with current awareness, but it should be relatively easy to encourage good habits at the start of their medical careers with the younger generation.

The incorporation of RSS into the medical syllabus at university would give all newly qualified medical professionals a basic understanding of what RSS is and its benefits, and could introduce them to a number of tools and services that would help them to better manage their data and keep abreast of the latest research without the daunting feeling of information overload. Employers and managers should also be educated as to the importance of RSS and requested to implement incentives to get their entire medical workforce familiar with its application. This may include pre-installing RSS software on all computers and handheld devices and ensuring that employees receive training and incentives for its use. As RSS is encouraged and adopted the benefits will become apparent to all.

In order to ensure that RSS updates are posted to the RSS feeds, the process must be included in the submission of any new data that is to be posted online.

6 Conclusion

RSS opens up a new dimension of information exchange that has the potential to enable giant steps forward in the field of medicine. Medical information published without the use of RSS may result in lost opportunities to quickly further

knowledge. For RSS to realise its full potential in the field of medicine, everyone needs to participate in the promotion of its use, by ensuring the provision of RSS feeds to announce the publication of all medically related data and the provision of RSS training wherever possible.

References

1. Glasziou, P.: Information overload: whats behind it, whats beyond it?, http://www.mja.com.au/public/issues/189_02_210708/gla10552_fm.html
2. Greenhalgh, T.: A comparative case study of two models of a clinical informaticist service. Volume 324., http://www.bmj.com/cgi/content/full/324/7336/524 (2002) 524–29
3. Hart, L.: Library 2.0: Rss feeds dynamic uses for special libraries., http://www.sla.org/pdfs/sla2007/hartrssfeeds.pdf
4. MedWorm: http://www.medworm.com
5. Wikipedia: Rss, http://en.wikipedia.org/wiki/RSS
6. Winner, D.: Rss 2.0 at harvard law., http://cyber.law.harvard.edu/rss/rss.html
7. Nottingham, M., Sayer, R.: Rfc 4278: The atom syndication format, http://tools.ietf.org/html/rfc4287 (Dec 2005)
8. NLM: Pubmed, http://www.ncbi.nlm.nih.gov/pubmed
9. WISC: Ebling library, http://ebling.library.wisc.edu/rss/index.cfm
10. WebMD: Medscape, http://www.medscape.com/
11. WebMD: http://www.webmd.com/
12. MediLexicon: Medical news today, http://www.medicalnewstoday.com/
13. Reuters: Reuters: Healthcare, http://www.reuters.com/sectors/healthcare
14. FDA: Fda: Rss feeds, http://www.fda.gov/AboutFDA/ContactFDA/StayInformed/RSSFeeds/default.htm
15. WHO: Who: Rss feeds., http://www.who.int/about/licensing/rss/en/
16. US_government: Flu.gov., http://www.pandemicflu.gov/
17. CDC: Cdc: Rss feeds, http://www2c.cdc.gov/podcasts/rss.asp
18. NHS: Nhs library, http://www.library.nhs.uk/
19. Powers, A.: Duke university medical library: Rss, http://www.mclibrary.duke.edu/training/rss
20. : The university of helsinki: Feednavigator, http://www.terkko.helsinki.fi/feednavigator/
21. Wittenbrink, H.: Rss and atom: Understanding and implementing content feeds and syndication
22. Hayes, B.: Spam, spam, spam, lovely spam. Volume 91(3)., http://www.americanscientist.org/issues/pub/2003/5/spam-spam-spam-lovely-spam (May-June 2003)
23. ClinicalTrials.gov: http://www.clinicaltrials.gov
24. RSS_Advisory_Board: Rss autodiscovery, http://www.rssboard.org/rss-autodiscovery (Nov 2006)
25. OPML: Userland software. opml 2.0, http://www.opml.org/spec2 (Nov 2007)
26. Zarodwny, J.: Rss auto-discovery 2.0, http://jeremy.zawodny.com/blog/archives/000967.html (Sep 2003)

Advantages	Disadvantages
A simple, accepted, widely used standard defining basic elements so that data can be easily shared across different platforms.	Defines only basic elements, so more complex data structures may be lost across platforms. Mal-formatted data and non-standard characters often cause RSS parsing to break.
It acts as an alert to the latest updates, saving the need to crawl the whole web to find the latest information.	Not all websites use RSS and others do not use it effectively. It depends on RSS items being posted to RSS feeds when new information is published.
The simple data structure removes problems related to redundant and poor quality data that is often added for the search engine optimization purposes.	Usually not all data is posted to a feed item, so a search on feed item content may miss important information and keywords are not always included. It does not therefore replace the need for a web crawler.
It can save bandwidth and speed up processing times by passing basic information and summaries of data rather than heavily formatted full text webpages.	It can result in an additional processing burden on a web server when some of the RSS feeds become heavily subscribed to, especially by other websites that may call the RSS feeds very frequently.
The manual collection of RSS feeds ensures the use of only quality data sources.	It is a manual process to find and add newly created RSS feeds to aggregators.
It gives the subscribers more privacy without the use of tracking cookies and scripts to log reader activity.	Readership of RSS items is difficult to track, due to the many different platforms through which data can pass and the lack of cookies and data tracking scripts.
It is known as 'pull' technology which allows the subscriber to decide what they wish to read and to have full control of which sources they no longer wish to subscribe to, hence the avoidance of spam.	RSS is not a push technology, which means that important updates can be missed if users have not subscribed.
The use of RSS encourages publishers to be more generous in the sharing of their data and hence promotes the advance of universal knowledge.	There is uncertainty and concern over the sharing of potentially copyrighted information via RSS.
Due to the passing of simply formatted text, RSS reader enable the user to scan through large amounts of data very quickly.	Formatting may be lost when data is passed through an RSS feed.
It can save RSS subscribers time and help keep them better informed on a topics of their interest.	It is hard be hard to motivate users to learn to use an RSS reader in addition to the use of email and there is the perceived risk of information overload.

Table 1. Advantages and Disadvantages of RSS

Animal Disease Event Recognition and Classification

Svitlana Volkova, Doina Caragea, William H. Hsu, Swathi Bujuru

Department of Computing and Information Sciences
Kansas state University, 234 Nichols Hall, Manhattan, KS, USA 66506
{svitlana,dcaragea,bhsu,swathi}@ksu.edu

Abstract. Monitoring epidemic crises, caused by rapid spread of infectious animal diseases, can be facilitated by the plethora of information about disease-related events that is available online. Therefore, the ability to use this information to perform domain-specific entity recognition and event-related sentence classification, which in turn can support time and space visualization of automatically extracted events, is highly desirable. Towards this goal, we present a rule-based approach to the problem of extracting animal disease-related events from web documents. Our approach relies on the recognition of structured entity tuples, consisting of attributes, which describe events related to animal diseases. The event attributes that we consider include animal diseases, dates, species and geo-referenced locations. We perform disease names and species recognition using an automatically-constructed ontology, dates are extracted using regular expressions, while location are extracted using a conditional random fields tool. The extracted events are further classified as confirmed or suspected based on semantic features, obtained from the *e.g., GoogleSets*[1] and *WordNet*[2]. Our preliminary results demonstrate the feasibility of the proposed approach.

Key words: entity recognition, animal disease, event tuple detection, classification, text mining

1 Introduction

The large spread of infectious diseases has a great negative impact on society. While human infectious diseases can result in significant loss of life, animal diseases can cause major problems across the world because of the influence on the economy and trade. Moreover, animal diseases that are zoonotic in type can also cause loss of life in addition to economic crises and political instability.

Infectious Disease Informatics (IDI) includes tasks such as: data collection, sharing, management, modeling and analysis in the domain of emerging infectious diseases [1]. An enormous amount of data about animal infectious disease-related events is available online in both structured and unstructured formats.

[1] GoogleSets Inteface - http://labs.google.com/sets
[2] WordNet - http://wordnet.princeton.edu/

Structured data is presented to public in official reports by different organizations such as: state and federal laboratories, local health care providers, governmental agricultural or environmental agencies. In addition, a lot of unstructured information can be found in a variety of other contexts *e.g.,* news, e-mails, blogs, which in contrast to the official reports is completely unorganized. In order to exploit this unstructured data, machine learning and text mining techniques can be used to recognize disease-related events, *e.g., "On 12 September 2007, a new foot-and-mouth disease outbreak was confirmed in Egham, Surrey"*. Such techniques could be part of automated systems that can detect, monitor and track responses to animal infectious disease outbreaks (defined as a set of events which are constrained in space and have temporal overlap).

Several automated systems for animal disease monitoring exist [2], [3]. They have the ability to crawl news and use ontology pattern matching approaches to recognize entities such as disease and location of an event. While the existing systems focus on news data and identify emergent diseases, in this paper we describe a system which can be used not only with news data, but also with e-mails, blog posts and scientific web articles. Therefore, our system can identify events in historical data as opposed to identifying only emergent disease events. Specifically, our system extracts event tuples from a variety of web documents. These tuples can be seen as structured summaries of the events specified by attributes such as: disease, location, date, species and confirmation status.

2 Methodology

In this section we describe in detail our methodology for identifying disease-related events and their associated confirmation status. The confirmation status refers to an event being suspected or confirmed. This information is important with respect to the action that needs to be taken. Our approach to the event recognition problem involves three main steps: first, we perform entity recognition from unstructured sources; next, we classify the sentences from which entities are extracted as being related to an event or not; furthermore, if they are related to an event we classify them as confirmed or suspected; finally, we combine entities within an event sentence into structured tuples. Figure 1 illustrates these three steps through an example.

2.1 Entity Recognition

The entity recognition module in our system automatically extracts structured information related to animal diseases from unstructured web documents. To achieve this functionality we associate meta-data in the form of ontologies with documents in our collection. Specifically, the meta-data consists of *domain-independent* location and time hierarchies (including names of countries, states, cities; and canonical dates) and a *domain-specific* medical ontology (including diseases, serotypes, and viruses). Based on these ontologies and pattern matching, we design specialized extractors that locate and classify atomic elements into predefined categories such as:

– disease names (*e.g., "foot and mouth disease", "rift valley fever")*;
– viruses (*e.g., "picornavirus")* and serotypes (*e.g., "Asia-1")*;
– species (*e.g., "sheep", "pigs", "cattle")*;
– locations of events specified at different levels of geo-granularity (*e.g., "United Kingdom", "eastern provinces of Shandong and Jiangsu, China")*;
– dates in different formats (*e.g., "last Tuesday", "two month ago").*

For the animal disease name recognition, we developed an Animal Disease Extractor (DSEx)[3], which relies on a medical ontology, automatically-enriched with synonyms and causative viruses [4]. For species extraction we use pattern matching on a stemmed dictionary of animal names from Wikipedia[4]. Furthermore, we used the Stanford NER[5] tool (which uses conditional random fields) together with NGA GEOnet Names Database (GNS)[6] for location recognition and set of regular expressions for date/time extraction.

The top panel in Figure 1 shows a paragraph where entities recognized by our extractors are highlighted. As an example, the output from our entity recognition module for the sentence *"Taiwan's TVBS television station reports that agricultural authorities confirmed foot-and-mouth disease on a hog farm in Taoyuan"* is shown below:

– animal diseases - *"foot-and-mouth disease"* (recognized by the DSEx);
– locations - *"Taoyuan"* (recognized by the Location Extractor);
– species - *"hog"* (recognized by the Species Extractor).

2.2 Event Sentence Classification

After the entities are recognized in a document, we next extract sentences that contain such entities and classify them as corresponding to true events or false positive events. True events should include a disease name together with a disease-related verb. Furthermore, these events are classified as confirmed or suspected using the Confirmation Status Extractor. This extractor relies on a restricted list of verbs that suggest confirmed events (*e.g., happened*) or suspected events (*e.g., catch*) and their synonyms identified using *GoogleSets*[1] or *WordNet*[2] [5] . For example, the following sentence is classified as corresponding to a confirmed event: *"On 9 Jun 2009, the farm's owner reported symptoms of FMD in more than 30 hogs."*

The initial list of verbs consists of single word verbs (*e.g., kill*) and verb phrases (*e.g., strike out*). The first two columns in Table 1 show the number of initial verbs denoted as $IN\text{-}V$ and verb phrases denoted as $IN\text{-}VP$ for both suspected and confirmed categories. Columns 3 and 4 show similar numbers for the augmented list of verbs obtained using *GoogleSets*[1] (denoted by $GS\text{-}V$, $GS\text{-}VP$ respectively), while columns 5, 6 show these numbers for *WordNet*[2] (denoted by $WN\text{-}V$, $WN\text{-}VP$ respectively).

[3] KDD DSEx - `http://fingolfin.user.cis.ksu.edu:8080/diseaseextractor/`
[4] Species in Wikipedia - `http://en.wikipedia.org/wiki/List_of_animal_names`
[5] Stanford NER - `http://nlp.stanford.edu/ner/index.shtml`
[6] GNS - `http://earth-info.nga.mil/gns/html/`

Table 1: Statistics about the restricted list of verbs

Status	IN-V	IN-VP	CS-V	GS-VP	WN-V	WN-VP
Suspected	7	1	55	2	37	10
Confirmed	7	1	55	13	48	9

The list of verbs used to classify sentences as confirmed or suspected is also useful for eliminating frequent, but not event-related sentences such as: "*Foot and mouth disease is[V] a highly pathogenic animal disease*".

The second step in Figure 1 shows more examples of potential event-related sentences and their classification. We first classify sentences as event-related (corresponds to *"YES"*) or event non-related (corresponds to *"NO"*). We then classify event-related sentences as suspected or confirmed based on the restricted list of verbs and verb phrases represented in Table 1.

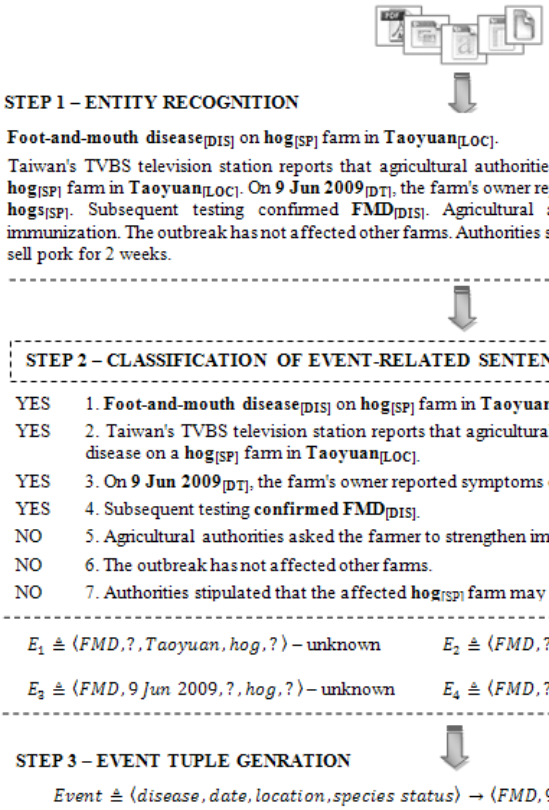


Fig. 1: Description of the system workflow through an example: first, entities are recognized using several extractors; second, the true event sentences are identified and classified as suspected or confirmed; next, instances from true event sentences are grouped together into potential event tuples; finally, instances of the same event are consolidated into one comprehensive tuple.

2.3 Event Tuple Generation

An *event* is an occurrence of a disease within a particular time and space range. We use four main event attributes to specify an event: disease name, date, location, species. In addition, as we extract events automatically from crawled web documents, we also include an attribute that specifies the confirmation status of an event. Thus, an event can be described as a tuple of the following form:

$$Event_i =< disease, date, location, species, status >, \quad (1)$$

where each attribute in the tuple obtained with one of the extractors described in Section 2.1. The following tuple <*FMD, 9 Jun 2009, Taoyuan, hog, confirmed*> is an example of an event. Given the incomplete and the uncertain nature of the information available online, it is possible for events to have missing values, *e.g.,* $< disease, ?, location, species, ? >=< FMD, ?, Taoyuan, hog, ? >$, $< disease, date, ?, species, ? >=< FMD$, *09 Jun 2009* $, ?, hog, ? >$. For instance, news reports can contain information about disease-related events that happened in some location without a specific date or species being provided.

Furthermore, several sentences in a document can contain information about the same event and we aggregate the corresponding event tuples into a unique tuple based on the attributes available, as shown in the last step in Figure 1.

Algorithm 1 Entity Recognition, Sentence Classification and Tuple Generation

```
Input: Set of web documents D
Output: Set of extracted events e_k ∈ E for each document d_j ∈ D
    foreach document d_j ∈ D do
     S = TokenizeToSentences(d_j);
       foreach sentence s_i ∈ S do
          disease = ExtractDiseaseEntities(s_i);
          if disease ≠ ∅ then
             status = ExtractConfirmationStatus(s_i);
             if status ≠ ∅ then
                date = ExtractDateEntities(s_i);
                location = ExtractLocationEntities(s_i);
                species = ExtractSpeciesEntites(s_i);
            else
                skip sentence s_i;
             end;
          else
             skip sentence s_i;
          end;
       end;
       E = GenerateTuples(disease, date, location, species, status);
       e_k = AggregateTuples(E);
    end.
```

Algorithm 1 summarizes the steps for entity recognition, event-related sentence classification and tuple generation.

3 Experimental Design and Results

We used the existing *DUCView Pyramid* scoring tool [6] to score automatically generated event tuples and evaluate our approach. Pyramid scoring is a technique for evaluating summarization results, which was introduced in [7] and relies on multiple summaries to assign the significance weights to summarization content units (*i.e.*, entities) [8].

To perform the evaluation, we used Google to retrieve 100 documents related to two animal diseases: rift valley fever (RVF) and foot-and-mouth disease (FMD). We manually created two sets of summaries for each of the 100 documents and extracted entities corresponding to event tuples from each summary and each document as described in Section 2.1. Then, we used the *DUCView* tool to compare automatically generated event tuples with entities from human summaries. As a result, the entities from event tuples are assigned weights in the range [0, 1] where 1 represents the best recognition score and it means that entity from automatically-generated tuple is present in all summaries. The entity weights are used to calculate an aggregated score for event tuples. Specifically, the score for an event tuple described in Equation 1 is given by:

$$Score_i =< w_d disease, w_t date, w_l location, w_s species, w_c status > \quad (2)$$
$$subject\ \ to\ \ disease + status = 2$$

where $disease, \cdots, species$ take 0/1 values (entity present or not in the tuple) and a tuple is valid only if both *disease* and *status* are present. The resulting scores are reported as a measure of the accuracy of the proposed event tuple recognition and classification approach and shown in Table 2.

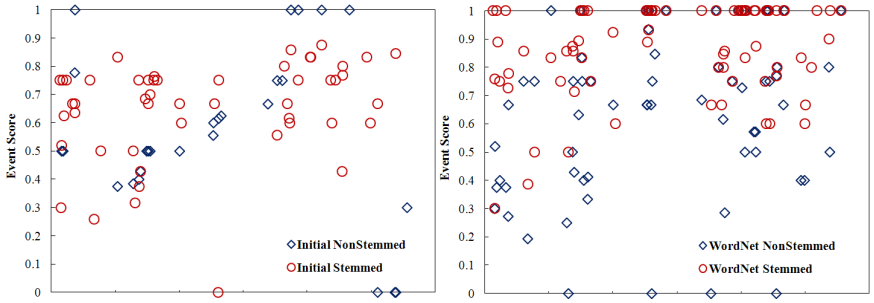
More precisely, we evaluate our event tuple recognition and classification approach by applying three lists of verbs and verb phrases for confirmation status extraction which are introduced in Table 1. Furthermore, we consider stemmed *S vs.* non-stemmed *NS* versions of these lists. The results for the non-stemmed version of the lists are shown in the first three columns of the Table 2 for the initial list, *GoogleSets*[1] augmented list and *WordNet*[2] augmented list, respectively. Similarly, the results for the stemmed version are shown in the last three columns of the Table 2.

Table 2: Pyramid Event Score Distribution by Range

Score Range	IN-NS	GS-NS	WN-NS	IN-S	GS-S	WN-S
Low [0 - 0.3]	73%	43%	38%	19%	18%	13%
Medium [0.31 - 0.7]	18%	27%	29%	27%	30%	13%
High [0.71 - 1]	9%	30%	33%	54%	52%	74%
Average Score	**0.17**	**0.40**	**0.45**	**0.64**	**0.65**	**0.75**

As can be seen from the Table 2, the initial list of verbs results in many low score events which means that not many tuples can be extracted with high confidence using only these verbs. While the augmented lists, without stemming, give better results, only approximately one third of the events are scored with a

high confidence for both *GoogleSets*[1] and *WordNet*[2]. However, the scores increase significantly for all lists when stemming is used. The best results are obtained for the *WordNet*[2] augmented list where the average score is as high as 0.75.



(a) Initial List (b) WordNet

Fig. 2: The scatter plot of the event scores using Pyramid method

Figure 2 shows the scatter plot of the event score distribution using the initial and *WordNet*[2] lists, for both stemmed and non-stemmed versions of lists. As can be seen, more events are identified using the *WordNet*[2] list and they have higher scores (many of them have the max score 1).

4 Related Work

There are several systems for disease-related event detection that extract diseases and locations from text. BioCaster[7] is an online ontology-based system for detecting and mapping infectious disease outbreaks from news [9]. Their approach for event detection is based on searching for disease-location pairs and calculating their frequency in the document and in the collection [2]. The methodology for deriving synonyms for disease-related verbs that are part of events (*e.g., disease, verb, location*) is similar to our approach. However, *BioCaster* does not provide assistance with classification of extracted events as confirmed or suspected. As opposed to *BioCaster, HealthMap*[8] is a manually supported web system, which crawls data from Google News and the ProMED-Mail[9] portal and provides reports about disease outbreaks to the public [10]. *Pattern-based Understanding and Learning System (PULS)*[10], which is part of the *MedISys*[11], allows extracting meta-data and structured facts related to the disease outbreaks [3]. Similar to other systems, it does not classify extracted events and does not report anything about past outbreaks. Our approach addresses the abovementioned limitations. It supports automated extraction of disease-related event tuples, which include disease, date, location, species entities and confirmation status. It also classifies them into two categories such as: suspected or confirmed.

[7] BioCaster Global Health Monitor - `http://biocaster.nii.ac.jp/`
[8] HealthMap System - `http://healthmap.org/en`
[9] ProMED-Mail - `www.promedmail.org`
[10] PULS - `http://sysdb.cs.helsinki.fi/puls/jrc/all`
[11] MedISys - `http://medusa.jrc.it/medisys/homeedition/all/home.htm`

5 Conclusions

In this paper, we presented an approach for animal disease event recognition and classification. Entity and confirmation status extraction methods are used to automatically generate structured summaries about domain-specific events in the form of tuples. Furthermore, we apply several lists of verbs for confirmation status extraction including *WordNet*[2] and *GoogleSets*[1]. We used the *Pyramid* method and *DUCView* tool [6] to calculate scores for automatically generated event tuples, which can be seen as a measure of accuracy of our approach. The highest accuracy was obtained using a *WordNet*[2] augmented list of verbs. As part of future work we intend to apply a deeper syntactic analysis of the sentence and part-of-speech tagging in addition to the list of verbs that we used.

Acknowledgments. This work was supported through a grant from the U.S. Department of Defense. We would like to acknowledge the Knowledge Discovery in Databases Laboratory assistants: John Drouhard, Landon Fowles (disease/species extractor), Wesam Elshamy, Andrew Berggren (location extractor), Danny Jones, Srinivas Reddy (date/time extractor).

References

1. Chen, H., Fuller, S.S., Friedman, C.P.: Medical Informatics: Knowledge Management and Data Mining in Biomedicine (Integrated Series in Information Systems). Springer (June 2005)
2. Kawazoe, A., Chanlekha, H., Shigematsu, M., Collier, N.: Structuring an event ontology for disease outbreak detection. BMC Bioinformatics **9 Suppl 3** (2008)
3. Steinberger, R., Fuart, F., Groot, E., Best, C., Etter, P., Yangarber, R.: Text mining from the web for medical intelligence. Mining Massive Data Sets for Security (2008)
4. Volkova, S., Hsu, W., Caragea, D.: Named entity recognition and tagging in the domain of epizootics (2009) Women in Machine Learning Workshop, `http://wimlworkshop.org/`.
5. Fellbaum, C.: WordNet: An Electronic Lexical Database. Bradford Books (1998) http://wordnet.princeton.edu/.
6. Nenkova, A.: Pyramid Annotation Guide - DUC 2006 `http://www1.cs.columbia.edu/~becky/DUC2006/2006-pyramid-guidelines.html_ducview`.
7. Nenkova, A.: Understanding the process of multi-document summarization: content selection, rewriting and evaluation. PhD thesis, New York, NY, USA (2006)
8. Nenkova, A., Passonneau, R., McKeown, K.: The pyramid method: Incorporating human content selection variation in summarization evaluation. ACM Trans. Speech Lang. Process. **4**(2) (2007)
9. Doan, S., QuocHung-Ngo, Kawazoe, A., Collier, N.: Global Health Monitor - a web-based system for detecting and mapping infectious diseases. In: Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP). (2008) 951–956
10. Freifeld, C., Mandl, K.D., Reis, B.Y., Brownstein, J.S.: Healthmap: Global infectious disease monitoring through automated classification and visualization of internet media reports. J Am Med Inform Assoc (December 2007)