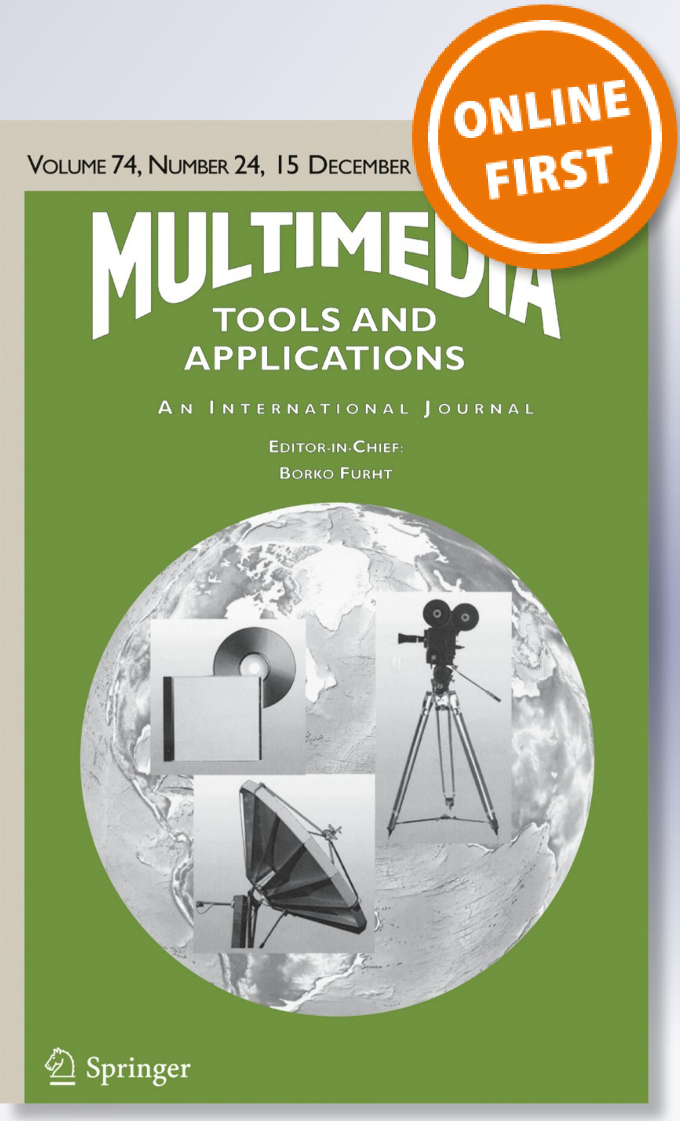
Topic categorization and representation of health community generated data

Maofu Liu, He Zhang, Huijun Hu & Wei Wei

Multimedia Tools and Applications
An International Journal

ISSN 1380-7501

Multimed Tools Appl
DOI 10.1007/s11042-015-3094-3



Multimed Tools Appl
DOI 10.1007/s11042-015-3094-3

Topic categorization and representation of health community generated data

Maofu Liu[1,2] · He Zhang[1,2] · Huijun Hu[1,2] · Wei Wei[3]

Received: 7 August 2015 / Revised: 14 October 2015 / Accepted: 17 November 2015

Abstract The representation and categorization of professional health provider released data have been well investigated and practically implemented. These have facilitated browsing, search and high-order learning of health information. On the other hand, there has been little corresponding studies on the representation and categorization of health community generated data. It is usually more complex, inconsistent and ambiguous, and consequently raises challenges for data access and analytics. This paper explores various representations for health community generated data and categorizes these data in terms of health topics. In addition, this work utilizes pseudo-labeled data to train the supervised topic categorization models, and this makes the whole categorization process unsupervised and extendable to handle large-scale data. The extensive experiments on two real-world datasets reveal our interesting findings of the informative representation approaches and effective categorization models for health community generated data.

Keywords Health community generated data · Learning model · Semantic representation · Health topic categorization

✉ Maofu Liu
e_mfliu@163.com

He Zhang
cheesezh@qq.com

Huijun Hu
huhuijun@wust.edu.cn

Wei Wei
weiw@hust.edu.cn

[1] College of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430065, China

[2] Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System, Wuhan 430065, China

[3] School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China

Published online: 24 November 2015

1 Introduction

Shifting demographics and increasing the cost of healthcare have driven consumers to explore health and wellness information online, which has reached almost saturation among American Internet users[1]. To better cater to the health seekers, numerous unprecedented online health services are springing up, such as WebMD[2], HealthTap[3] and PatientsLikeMe[4]. They provide trustworthy health knowledge, supportive community, timely health news, and other services.

Generally speaking, the current online health services can be roughly classified into two sources. The first one includes those released by professional health providers, such as WebMD, Yahoo! Health[5] and Drugs[6]. They are usually well-written, authoritative, comprehensive and organized. The other source is the community generated information, such as HealthTap, PatientsLikeMe and HaoDF[7]. They are blending content of patient asked questions, history records, expert commentaries and medical reviews. Compared to the former one, they are usually unstructured, typically presented in narrative language, frequently update, and do not follow standard naming conventions. In addition, there exists abundant spelling errors and acronyms with multiple possible meanings. However, we firmly believe that these two sources of health knowledge both hold great potential to support health information seeking and decision making surrounding health education and self-care.

In fact, the representation and categorization of the first source have attracted worldwide attentions in both academic and industrial circles for a long time [10]. For example, the previous efforts [20, 21] have demonstrated that the search and learning performance in terms of medical concept-based representation [5] can be inconsistent, sometimes underperforming the traditional term-based representation. This finding dramatically boosts the performance of health search and learning algorithms [11, 36]. More importantly, health provider released data has been practically categorized into health topics in plat structure, as shown in Fig. 1. This categorization method enables browsing and navigational style of information access.

On the other hand, representation and categorization of health community generated data seem more urgent due to the accumulated mountains of data in a faster pace. The corresponding studies, however, are still relatively sparse. To bridge this research gap, in this paper, we explore various representation approaches and utilize these representations to aid categorization. To be more specific, we first investigate traditional term-based, domain specific and Latent Dirichlet Allocation (LDA) based semantic representations respectively. We then feed these representations into widely-adopted supervised categorization models to predict the health topics. Health topic is a medical concept that can semantically summarize the given data instances from a high-level angle, which can be symptom, treatment or disease concept. It is worth mentioning that the supervised categorization models are trained with pseudo-labeled samples. Therefore, the whole process is automatic and applicable to handle large-scale data. Extensive experimental results on two real-world datasets answer these two questions, which representation and model should be chosen to categorize the health community generated data.

[1] http://pewinternet.org/Reports/2013/Health-online.aspx

[2] www.webmd.com

[3] https://www.healthtap.com

[4] www.patientslikeme.com

[5] http://health.yahoo.net

[6] www.drugs.com

[7] www.haodf.com

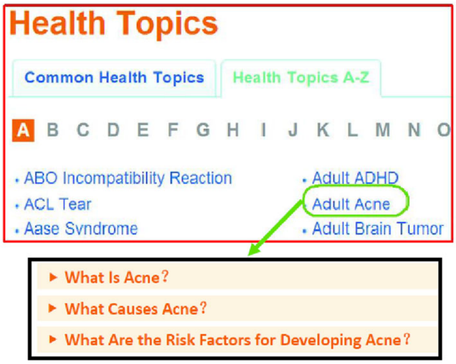


Fig. 1 Categorization illustration of health provider released data. This sample is selected from Yahoo! Health

The main contributions of this paper are threefold:

(1) To the best of our knowledge, this is the first work on automatically categorizing health community generated data in terms of health topics. That is fundamental to make data more accessible.
(2) To support topic categorization, we explore and exploit various representations for health community generated data, spanning from low-level syntactic to high-level semantic.
(3) We evaluate several state-of-the-art supervised learning models on two pseudo-labeled datasets and find that Support Vector Machine (SVM) shows superiority over other approaches for the topic categorization of health community generated content.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 introduces the topic categorization and representations of health community generated data. Section 4 then presents experimental results and discussions. Finally, Section 5 concludes the paper.

2 Related work

The healthcare information processing has attracted more research attentions in recent years. Srinivasan [20] offered a detailed analysis, indicating that there are a number of open questions relevant to the overall debate on indexing vocabularies for MEDLINE, and eight different retrieval strategies involving document indexing via free-text, MeSH and several alternative combinations of the two vocabularies. Yang et al. [28] analyzed geotagged mobile queries in a privacy-sensitive study of potential transitions from health information search to in-world healthcare utilization. Nie et al. [15] reported a user study on the information needs of health seekers in terms of questions, especially those asking for possible diseases of their manifested symptoms, and proposed a novel deep learning scheme to infer the possible diseases given the questions of health seekers. Nie et al. [13, 16] also presented the WenZher system, which automatically organizes all the associated healthcare knowledge into a single view for a given question and generates the comprehensive answer from heterogeneous and multilingual data sources, to overcome the information over-loading. In this paper, we choose the health community generated data as the corpus, instead of those released by the professional health providers.

Hong et al. [6] initially investigated on multimedia question answering (MMQA) from large scale data and Nie et al. [13] proposed the MMQA approach to enrich community-contributed textual answers in community Question Answering (cQA) with appropriate multimedia data for more complex questions. Qu et al. [19] studied the problem of question topic classification using a very large real-world cQA dataset, comprising 3.9 million questions and organizing into more than 1000 categories in a hierarchy, from Yahoo! Answers. Chan et al. [3] presented a hierarchical kernelized model for the automatic classification of general questions into their corresponding topic categories in cQA services. Kanavos et al. [7] proposed a novel system to address the issues of clustering biomedical search engine results according to their topic. In this paper, we try to classify the health community generated data into the health topics.

Machine learning intends to make computer simulate and evolve human behaviors based on different types of empirical data, and the machine learning methods can be divided into several types according to their mechanism, including supervised, semi-supervised, weakly supervised, unsupervised and so on. The machine learning methods can be used to solve the classification problem [30, 34]. Yan et al. [23] proposed the multi-task learning framework for classifying the head poses of a person. Chang et al. [4] defined the notion of semantic saliency assessing the relevance of each shot with the event of interest and proposed the nearly-isotonic SVM classifier to discriminate the semantic ordering information. The weakly supervised image segmentation model focusing on learning the semantic associations in the image was put forward by Zhang et al. [31]. Yan et al. [26] proposed the multi-task unsupervised clustering framework for the activity of daily living analysis. In this paper, we evaluate several state-of-the-art supervised learning models on two pseudo-labeled health community generated datasets.

Limsopatham et al. [10] put forward a novel technique to represent medical records and queries by focusing only on medical concepts essential for the information need of a medical search task. In another paper, they [11] implemented a learning framework to model the importance of the bag-of-words and the bag-of-concepts representations, combining their scores on a per-query basis. Zhang et al. [29] presented a novel term weighting scheme with question retrieval models by incorporating the dependency relation cues between term pairs in cQA services. Zhu et al. [36] proposed a query-adaptive weighting method that can dynamically aggregate and score evidence in multiple medical reports of a patient. Trieschnigg et al. [21] approached the incorporation of a concept-based representation in the monolingual biomedical information retrieval from a cross-lingual perspective and realized it by translating and matching between text and concept-based representations. Babashzadeh et al. [1] attempted to use semantic information to improve the performance of clinical information retrieval systems by representing queries in an expressive and meaningful context. Li et al. [9] proposed a novel semantic-based approach to achieving the diversity-aware retrieval of Electronic Medical Records (EMRs) by exploiting the uncertainty in ambiguous medical queries. Nie et al. [17, 18] presented the scheme jointly utilizing local mining and global learning approaches to bridging the gaps among health seekers, providers and community generated knowledge. In fact, some methods and models of image semantic representation or feature selection can also be referred to for the health community generated data [32, 35]. Zhang et al. [33] presented a weakly-supervised image segmentation algorithm that learns the distribution of spatially structural superpixel sets from image-level labels. Yan et al. [27] introduced the event oriented dictionary representation based on the selected semantic meaningful concepts. In their other two papers, they [24, 25] explored the feature selection with sparsity coupling GLocal structural for multimedia classification and understanding. In this paper, we investigate on the term-based, domain specific and LDA-based topic representation of the health community generated data.

3 Topic categorization and representations

This section introduces topic categorization and various kinds of topic representations for the health community generated data.

3.1 Categorization

As mentioned previously, the health provider released data is organized in terms of health topics, which are sorted and displayed alphabetically. According to our statistics, the number of well-known topics is approximately one thousand and a half. In this paper, we work towards organizing the health community generated data by assigning each data instance into one of these health topics. We view this assignment problem as a categorization task, where each health topic is assumed to be one category. Four prevailing categorization models are employed to accomplish this task, i.e. Naïve Bayes, K Nearest Neighbor (KNN), Decision Tree C4.5 and SVM. The strongest model will stand out via comparative evaluation.

To comprehensively validate these models, two groups of health topics were selected from WebMD and listed in Table 1. The topics in the first group were randomly selected. While those in the second group were manually and deliberately chosen to ensure they are semantically and syntactically similar. This procedure makes the topic discrimination harder and can validate the robustness of the selected models.

These topics are naturally regarded as queries. Moreover, two datasets were constructed by retrieving all the related question answering (QA) pairs from HealthTap based on these two groups of queries. To avoid the expensive manual labeling procedure, the returned QA pairs were directly considered as the positive samples under associated topics. This is the so-called pseudo labeled ground truth. It makes the whole categorization process unsupervised and extendable to handle large-scale data.

Table 1 The illustration of selected health topics

ID	First Group	Second Group
1	Aase Syndrome	Temporal Lobe Epilepsy
2	ABO Incompatibility Reaction	Temporomandibular Joint Syndrome
3	Bacterial Arthritis	Diabetes, Gestational
4	Bleeding Esophageal Varices	Diabetic Ketoacidosis
5	Breast Cancer	Diabetic Neuropathy
6	Brittle Bone Disease	Gingivitis
7	Canker Sore	Teething
8	Closed Angle Glaucoma	Tooth Decay
9	Corns and Callosities	Breast Lump
10	Deep Vein Thrombosis	Breast Milk Jaundice

3.2 Term-based representation

This approach represents each QA pair as vectorized collection of term-based bags, disregarding grammar and even word order but keeping multiplicity. It is commonly used in document categorizations, where the occurrence of each word is used as a feature for training a classifier.

In our work, 20,619 and 19,193 distinct terms were respectively identified from the first and second datasets. The terms with frequencies smaller than six were empirically removed. It is worth noting that for the stop words removal, the interrogative terms such as what and why, were kept, since they play essential roles in interpreting the question semantics [8, 12]. We also did some simple preprocessing to link the variants together, such as singularizing plural variants. Ultimately, we obtained 5036 and 4757 dimensional term-based features for the first and second dataset, respectively.

3.3 Domain specific representation

We may argue that the health domain specific concepts may be more descriptive and capable of capturing the characteristics of health sources [8, 14]. To gather the health concepts, we first extract all the embedded noun phrases from corpus and then identify the health concepts from these noun phrases by measuring their specificity.

Initially, we assign part-of-speech (POS) tags to each word in the given corpus by Stanford POS tagger[8]. We then identify the sequences that match a fixed pattern of POS tags as noun phrases. The pattern is formulated as,

$$(Adjective|Noun)^{*}(Noun\ \ Preposition)?(Adjective|Noun)^{*}Noun \tag{1}$$

A sequence of tags matching this pattern ensures that the corresponding text fragment make up a noun phrase.

To differentiate the health concepts from other general noun phrases, we assume that concepts that are relevant to health domain frequently occur in health specific corpus and rarely in non-health one [22]. Based on this assumption, we employ the concept entropy impurity (CEI) [28] to comparatively measure the domain-relevance of a concept. For a concept c, its CEI is computed as follows,

$$\mathrm{CEI}(c) = -\sum_{i=1}^{2} P(D_i|c)\log P(D_i|c) \tag{2}$$

where D_1 and D_2 respectively represents our health corpus and a general domain corpus[9], and $P(D_i|\,c)$ denotes the probability that a concept c is related to a specified domain D_i. To make it easily computer-processable, we define specificity of a concept to the health domain as follows,

$$specificity(c) = \begin{cases} 1-\alpha CEI(\mathrm{c}) & if \quad P_n(D_1|c) > P_n(D_2|c) \\ \alpha CEI(\mathrm{c}) & otherwise \end{cases} \tag{3}$$

where $\alpha = {}^{0.5}\!/_{0.693}$. Meanwhile, a threshold is set to detect the health concepts.

[8] http://nlp.stanford.edu/software/tagger.shtml

[9] In this work, D_2 is a general English Gigaword data of Linguistic Data Consortium (http://www.ldc.upenn.edu/)

We separately detected 1779 and 1419 distinct health concepts for two datasets. Moreover, each QA pair will be represented by health concept histogram. The health concept frequency distribution in the first dataset is illustrated in Fig. 2.

From this figure, it is observed that the health concepts with higher frequencies are usually more generic and less informative, such as "women's health" and "pain". While health concepts with rare occurrences are very specific and descriptive, such as "rotavirus infection" and "muscle paralysis". Hence, it is unreasonable to treat each dimension equally. To address this problem, we adaptively weight each health concept in terms of its frequency,

$$r(c) = \frac{1}{\log(o(c) + 1)} \tag{4}$$

where $o(c)$ refers to the occurrence frequency of health concept c. This formula stamps the generic health concepts and rewards the specific health concepts. It can substantially strengthen the representativeness of each health concept. Finally, each feature vector is normalized to have zero-mean and one-variance.

3.4 LDA-based semantic representation

The term-based and domain specific representation approaches for health data typically lead to an explosion of feature dimension. In addition, they are unable to recognize synonyms from a given term set or concept set, let alone to recognize semantic relationships between terms or concepts.

In this work, we explore a high-level feature, the LDA-based semantic representation. It is able to describe the underlying semantic structures of health data. LDA [2] models every

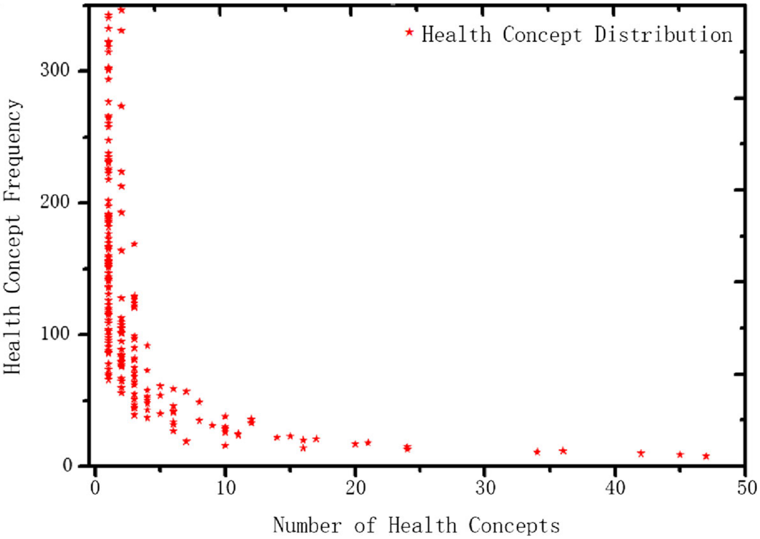


Fig. 2 The distribution of health concept frequency over the number of distinct health concepts

unobserved group as a distribution over the words of the vocabulary, and every observed document as a distribution over the groups. Here words assigned into the same group are semantically inter-related. We can thus use the latent group mixture of a document as a reduced representation.

To determine the number of latent groups, perplexity is adopted. Perplexity is commonly used as a quantitative measure of the coverage of the language model. A lower perplexity score indicates better-generalized performance. For a set of M documents, it is typically formulated as,

$$Perplexity(D) = \exp\left\{\frac{\sum_{d=1}^{M} \log(w_d)}{\sum_{d=1}^{M} N_d}\right\} \tag{5}$$

The Stanford topic modeling toolbox[10] was employed to model our two datasets, respectively. The perplexity curve of the first dataset is displayed in Fig. 3. As illustrated, the perplexity curve sharply goes down with group number growing and arrives at a trough at a certain group number, and the perplexity then goes upward and finally becomes relatively stable. From this table, it can be seen that 220 dimensional LDA-based semantic features can better capture our first dataset. Similarly, the second dataset is quantized and vectorized by 100 dimensional LDA-based features.

4 Experimental results

With topics as queries, we crawled community generated data from HealthTap, which is a silicon valley-based QA site for consumers to ask health-related questions, and receive answers with tags from certified physicians. Two datasets were constructed with two groups of health topics, and they respectively include 5076 and 4952 data instances. Each data instance includes three kinds of information cues, i.e. question, answers and multiple tags associated with answers.

Each dataset was divided into two subsets, i.e. 80 % as the training set and 20 % as the testing set. Four classifiers were respectively trained on the training set with various representations and validated on the testing set. Table 2 shows the comparative evaluation results on the first dataset of four different models with three different features. And the results on the second dataset are displayed in Table 3. Overall, they all achieved promising performance. Joint analytics of these two tables, we can see that the best performance on the first dataset slightly outperforms that on the second dataset. This is due to the different inherent data structures: the topics for the first dataset are more distinguishable, while those in the second dataset are similar which may share overlapped terms, concepts and semantic meanings.

Overall, the idea that automatically categorizing health community generated data in terms of health topics is reasonable.

[10] http://nlp.stanford.edu/downloads/tmt/tmt-0.4/

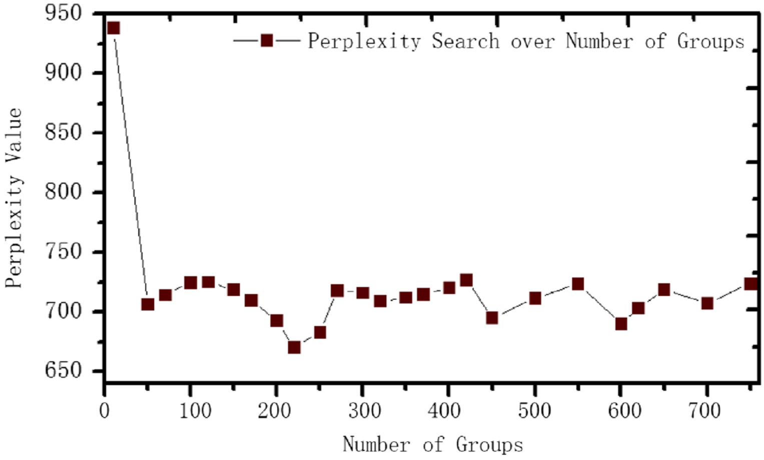


Fig. 3 The variation of perplexity value with the number of groups

Among these four categorization models, SVM shows superiority over other approaches. And the overarching trend is the performance goes up when more information cues are involved in training the models. From these two tables, we can infer that the verbose answers contain more descriptive information compared to the short questions, and they play essential roles in boosting the categorization performance. We can also observe that the manually annotated tags are able to significantly advance accuracies of models.

For the representation approaches, we have one interesting finding: Under the same categorization models, the performance of term-based approach markedly and significantly outperforms LDA-based semantic representation and is consistently better than domain specific representation. One possible reason is the information loss. LDA-based and domain specific representation is able to respectively seize the high-level semantic grouping information and domain-knowledge interpretations, while they overlook some specific terms. These specific terms, even not primary factors, may have some effects in the health contextualization. However, as compared to term-based approach, the dimension respectively increases more than twenty-fold for LDA-based and threefold for domain specific. These ratio values will grow continuously as the dataset size increases. Therefore, under certain circumstances, such as large-scale datasets, LDA-based and domain specific representations may be better alternatives.

Table 2 The comparative evaluation results on the first dataset of four different models with three different representations. Here Q, Q + A and Q + A + T respectively denotes question, QA pair, and QA pair with tags

Models	Term-based			Domain Specific			LDA-based		
	Q	Q + A	Q + A + T	Q	Q + A	Q + A + T	Q	Q + A	Q + A + T
Naïve Bayes	93.3 %	93 %	99.5 %	77.4 %	73.4 %	81.8 %	54.4 %	69.8 %	75.8 %
KNN	87.8 %	65.7 %	96.4 %	75.8 %	68.8 %	81.2 %	60.2 %	75 %	73.4 %
Decision Tree	92.7 %	95.4 %	99.9 %	76.3 %	83.8 %	94 %	81.2 %	82.2 %	88.1 %
SVM	94.1 %	96.4 %	99.9 %	81.1 %	83.4 %	93 %	80.3 %	85.9 %	87.8 %

Table 3 The comparative evaluation results on second dataset of four different models with three different representations. Here Q, Q + A and Q + A + T respectively denote question, QA pair, and QA pair with tags

Models	Term-based			Domain Specific			LDA-based		
	Q	Q + A	Q + A + T	Q	Q + A	Q + A + T	Q	Q + A	Q + A + T
Naïve Bayes	93.6 %	91.5 %	96.9 %	67.8 %	74.6 %	84.4 %	70.1 %	78 %	80.3 %
KNN	91.4 %	61.7 %	89.2 %	66.2 %	72 %	81.1 %	61 %	76.4 %	85.4 %
Decision Tree	94.3 %	95.6 %	99.2 %	65.9 %	87.4 %	96 %	77.2 %	80 %	84.6 %
SVM	95 %	95.8 %	99.3 %	69.3 %	83.9 %	94.2 %	83.2 %	86.4 %	92.8 %

5 Conclusions

This paper addressed the categorization and representation problems in community-based health services. This is a complementary work to the studies on professional health provider released data. We viewed data organization as a topic categorization problem. Several award-winning supervised learning models with various data representations were evaluated on two pseudo-labeled datasets. We observed that the SVM model with term-based representation achieves the best performance. While the term-based approach has high feature space as compared to other two approaches.

Acknowledgments The work presented in this paper is partially supported by the National Natural Science Foundation of China under Grant No. 61100133 and the Major Projects of National Social Science Foundation of China under Grant No. 11&ZD189.

References

1. Babashzadeh A, Huang J, Daoud M (2013) Exploiting semantics for improving clinical information retrieval. Proceedings of the International ACM SIGIR Conference 801–804
2. Blei D, Ng A, Jordan M, Lafferty J (2003) Latent dirichlet allocation. J Mach Learn Res 3:993–1022
3. Chan W, Yang W, Tang J, et al (2013) Community question topic categorization via hierarchical kernelized classification. Proceedings of the 22nd ACM International Conference on Information and Knowledge Management 959–968
4. Chang X, Yang Y, Xing E, Yu Y (2015) Complex event detection using semantic saliency and nearly-isotonic SVM. Proceedings of the 32nd International Conference on Machine Learning 1348–1357
5. Hersh W, Hickam D, Haynes R, Mckibbon K (1994) A performance and failure analysis of SAPHIRE with a MEDLINE test collection. J Am Med Inform Assoc 1(1):51–60
6. Hong R, Li G, Nie L, Tang J, Chua T (2010) Exploring large scale data for multimedia QA: an initial study. Proceedings of the ACM International Conference on Image and Video Retrieval 74–81
7. Kanavos A, Makris C, Theodoridis E (2015) Topic categorization of biomedical abstracts. Int J Artif Intell Tools. doi:10.1142/S0218213015400047
8. Kim M and Goebel R (2010) Detection and normalization of medical terms using domain-specific term frequency and adaptive ranking. IEEE International Conference on Information Technology and Applications in Biomedicine 1–5
9. Li J, Liu C, Liu B, Mao R, Wang Y, Chen S, Yang J, Pan H, Wang Q (2015) Diversity-aware retrieval of medical records. Comput Ind 69:81–91
10. Limsopatham N, Macdonald C and Ounis I (2013a) A task-specific query and document representation for medical records search. Proceedings of the European Conference on Advances in Information Retrieval 747–751

11. Limsopatham N, Macdonald C and Ounis I (2013b) Learning to combine representations for medical records search. Proceedings of the International ACM SIGIR Conference 833–836
12. Nie L, Wang M, Zha Z, Li G, and Chua T (2011) Multimedia answering: Enriching text QA with media information. Proceedings of the International ACM SIGIR Conference 695–704
13. Nie L, Wang M, Gao Y, Zha Z, Chua T (2013a) Beyond text QA: multimedia answer generation by harvesting web information. IEEE Trans Multimedia 15(2):426–441
14. Nie L, ZhaoY WX, Shen J, Chua T (2013b) Learning to recommend descriptive tags for questions in social forums. ACM Trans Inf Syst 32(1):5. doi:10.1145/2559157
15. Nie L, Wang M, Zhang L, et al. (2014a) Disease inference from health-related questions via sparse deep learning. IEEE Trans Knowl Data Eng 27(8):2107–2119
16. Nie L, Li T, Akbari M, Shen J, Chua T (2014b) WenZher: comprehensive vertical search for healthcare domain. Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval 1245–1246
17. Nie L, Akbari M, Li T, Chua T (2014c) A joint local-global approach for medical terminology assignment. In Medical Information Retrieval Workshop at SIGIR 2014, 24–27
18. Nie L, Zhao Y, Akbari M, Shen J, Chua T (2015) Bridging the vocabulary gap between health seekers and healthcare knowledge. IEEE Trans Knowl Data Eng 27(2):396–409
19. Qu B, Cong G, Li C, et al. (2012) An evaluation of classification models for question topic categorization. J Am Soc Inf Sci Technol 63(5):889–903
20. Srinivasan P (1996) Optimal document-indexing vocabulary for MEDLINE. Inform Process Manag 32:503–514
21. Trieschnigg D, Hiemstra D, de Jong F and Kraaij W (2010) A cross-lingual framework for monolingual biomedical information retrieval. Proceedings of the ACM Conference on Information and Knowledge Management 169–178
22. Velardi P, Missikoff M and Basili R (2001) Identification of relevant terms to support the construction of domain ontologies. Proceedings of the workshop on Human Language Technology and Knowledge Management, doi:10.3115/1118220.1118225.
23. Yan Y, Ricci E, Subramanian R, Lanz O, Sebe N (2013a) No matter where you are: Flexible graph-guided multi-task learning for multi-view head pose classification under target motion. Proceedings of 2013 I.E. International Conference on Computer Vision 1177–1184
24. Yan Y, Xu Z, Liu G, Ma Z, Sebe N (2013b) GLocal structural feature selection with sparsity for multimedia data understanding, Proceedings of the ACM International Conference on Multimedia 537–540
25. Yan Y, Shen H, Liu G, Ma Z, Gao C, Sebe N (2014) GLocal tells you more: coupling GLocal structural for feature selection with sparsity for image and video classification. Comput Vis Image Underst 124: 99–109
26. Yan Y, Ricci E, Liu G, Sebe N (2015a) Egocentric daily activity recognition via multitask clustering. IEEE Trans Image Process 24(10):2984–2995
27. Yan Y, Yang Y, Meng D, Liu G, Tong W, Hauptmann A, Sebe N (2015b) Event oriented dictionary learning for complex event detection. IEEE Trans Image Process 24(6):1867–1878
28. Yang S, White R and Horvitz E (2013) Pursuing insights about healthcare utilization via geocoded search queries. Proceedings of the International ACM SIGIR Conference 993–996
29. Zhang W, Ming Z, Zhang Y, Nie L, Liu T, Chua T (2012) The use of dependency relation graph to enhance the term weighting in question retrieval. Proceedings of the 25th International Conference on Computational Linguistics 3105–3120
30. Zhang L, Han Y, Yang Y, Song M, Yan S, Tian Q (2013) Discovering discriminative graphlets for aerial image categories recognition. IEEE Trans Image Process 22(12):5071–5084
31. Zhang L, Yang Y, Gao Y, Yu Y, Wang C, Li X (2014a) A probabilistic associative model for segmenting weakly supervised images. IEEE Trans Image Process 23(9):4150–4159
32. Zhang L, Gao Y, Ji R, Xia Y, Dai Q, Li X (2014b) Actively learning human gaze shifting paths for semantics-aware photo cropping. IEEE Trans Image Process 23(5):2235–2245
33. Zhang L, Gao Y, Xia Y, Lu K, Shen J, Ji R (2014c) Representative discovery of structure cues for weakly-supervised image segmentation. IEEE Trans Multimedia 16(2):470–479
34. Zhang L, Gao Y, Xia Y, Dai Q, Li X (2015a) A fine-grained image categorization system by cellet-encoded spatial pyramid modeling. IEEE Trans Ind Electron 62(1):564–571
35. Zhang L, Xia Y, Mao K, Ma H, Shan Z (2015b) An effective video summarization framework toward handheld devices. IEEE Trans Ind Electron 62(2):1309–1316
36. Zhu D and Carterette B (2013) An adaptive evidence weighting method for medical record search. Proceedings of the International ACM SIGIR Conference 1025–1028

Maofu Liu is currently a Professor in College of Computer Science and Technology of Wuhan University of Science and Technology. He received his Ph.D, M.Sc and B.Sc degrees from Wuhan University in Computer Science in 2005, 2002 and 1998 respectively. His main research interests include natural language processing, image processing and machine learning.

He Zhang is currently M.Sc candidate in College of Computer Science and Technology of Wuhan University of Science and Technology. He received his B.Sc degree from College of Computer Science and Technology of Wuhan University of Science and Technology in 2013. His main research interests include natural language processing and machine learning.

Huijun Hu is a Ph.D. candidate in the State Key Lab of Software Engineering of Wuhan University and also a Lecturer in College of Computer Science and Technology of Wuhan University of Science and Technology. She received her M.Sc degree from Wuhan University in Computer Science in 2006. Her current main research interests include optimization theory, image processing and text processing.

Wei Wei is a Lecturer in School of Computer Science and Technology of Huazhong University of Science and Technology. He received his Ph.D, M.Sc and B.Sc degrees from Huazhong University of Science and Technology in Computer Science in 2012, 2008 and 2006 respectively. His current main research interests include information retrieval, data mining and natural language processing.