MEDINFO 2013
C.U. Lehmann et al. (Eds.)

doi:10.3233/978-1-61499-289-9-637

Automatically Identifying Health- and Clinical-Related Content in Wikipedia

Feifan Liu[a*], Soheil Moosavinasab[b*], Shashank Agarwal[c], Andrew S. Bennett[d,e], Hong Yu[f,g]

[a] *NLP R&D, Nuance Communication Inc., Burlington, MA, USA*
[b] *University of Wisconsin Milwaukee, Milwaukee, WI, USA*
[c] *DataXu Inc., Boston, MA, USA*
[d] *Medical College of Wisconsin, Milwaukee, WI;* [e] *Clement J Zablocki VA Medical Center, Milwaukee WI*
[f] *University of Massachusetts Medical School, Worcester, MA;* [g] *VA Central Western Massachusetts, Northampton, MA*

Abstract

Physicians are increasingly using the Internet for finding medical information related to patient care. Wikipedia is a valuable online medical resource to be integrated into existing clinical question answering (QA) systems. On the other hand, Wikipedia contains a full spectrum of world's knowledge and therefore comprises a large partition of non-health-related content, which makes disambiguation more challenging and consequently leads to large overhead for existing systems to effectively filter irrelevant information. To overcome this, we have developed both unsupervised and supervised approaches to identify health-related articles as well as clinically relevant articles. Furthermore, we explored novel features by extracting health related hierarchy from the Wikipedia category network, from which a variety of features were derived and evaluated. Our experiments show promising results and also demonstrate that employing the category hierarchy can effectively improve the system performance.

Keywords:

Natural Language Processing, Machine Learning.

Introduction

Physicians are increasingly using the Internet to find medical information related to patient care [1]. As a collaboratively written Web-based encyclopedia, Wikipedia [2] has evolved into an important medical resource for the general public, students, and healthcare professionals [3]. Studies show that 70% of junior physicians use Wikipedia during a given week, and nearly 50% to 70% of practicing physicians use it as an informational source while providing medical care (e.g. [4]). The junior physicians use Wikipedia more frequently than all other websites excluding Google [4]. Another study shows that in a survey, 35% of pharmacists admitted to using Wikipedia for medical information [5]. Among online health information resources, Wikipedia has shown to be a prominent source, ranking among the top 10 results in 71-85% of search engines and keywords tested [6]. The study also showed that Wikipedia surpassed MedlinePlus and NHS Direct Online, and ranked higher with quality articles.

Clinical Question Answering (QA) systems enable physicians to efficiently seek succinct and accurate answers to questions during their patient encounter. However, existing clinical QA systems, such as AskHERMES [7] and MiPACQ [8], are typically built upon secure knowledge resources (e.g. the biomedical literature), Wikipedia therefore has a potential to improve the QA performance by providing additional complementary and reliable resources.

On the other hand, Wikipedia contains a full-spectrum of global knowledge. The enormous quantity of non-health related information, if incorporated into clinical QA systems, could adversely affect the system's accuracy and efficiency. Firstly, there is a large amount of ambiguities between the health domain and other domains. Information retrieval components in clinical QA systems would be more susceptible to retrieving spurious documents from Wikipedia. Secondly, using category information assigned to Wikipedia articles to identify health related articles is possible; however, articles are often assigned categories from multiple domains, making the use of category information unreliable. Automatically identifying health related content and excluding non-health related content becomes a necessary step to incorporate Wikipedia knowledge into clinical QA systems. In addition, a health-related Wikipedia corpus could benefit other biomedical natural language processing tasks such as machine translation.

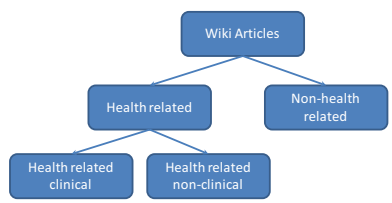

Figure 1 - Classification scheme illustration

In this study, we formulate this task into a traditional text classification task, where the goal is to predict the class[1] label for each Wikipedia article. Figure 1 shows the classification scheme defined for this task. At a high level, Wikipedia articles can be classified into "Health related" and "Non-health related." Those "health related" articles can be further classified into "health related clinical" and "health related non-clinical". A "health related clinical" article refers to articles with content pertaining directly to an aspect of a human disease or condition, such as causes, treatments, diagnosis, prognosis, therapeutic response, therapeutic drug levels, or clinical outcome. "Health related non-clinical" articles refer to those that are not clinical related but have an effect on conditions of human body or mind (e.g., wellness, healthcare organization). Articles not belonging to either of the two health-related classes are "Non-health related". The motive to further delineate the health related class into clinical versus non-clinical was that clinically related content is more beneficial for a clinical QA system.

We explored both unsupervised and supervised approaches for this task. In addition to traditional bag-of-words (BOW) linguistic features, we extracted health related category hierarchies from the Wikipedia category network. From the hierar-

* Co-First authors contribute equally.
1 Note that "category" denotes the subject information assigned to each Wikipedia article by Wikipedia editors; "class" for the classification scheme on which we formulated our task and build the model.

chies, a group of novel features were derived and evaluated. Our contributions are:

1. A pilot study that automatically identifies health related content from Wikipedia.
2. Exploitation of category hierarchies in Wikipedia to improve classification performance.
3. A first step towards optimally integrating reliable WWW resources into a clinical QA system.

Related Work

There has been increasing amounts of work done to mine Wikipedia knowledge for QA systems (e.g. [9]). However, most existing work focuses on open-domain factoid/definitional questions. To our best knowledge, no studies have explored integrating Wikipedia into a clinical QA system.

In the healthcare domain, Wikipedia has become an important medical resource [3]. Many studies have focused on evaluating content quality and coverage in Wikipedia. Friedlin et al. [10] concluded that Wikipedia contains a 'surprisingly' large amount of scientific and medical data that could effectively be used as a knowledge base for specific medical informatics or research projects. Rajagopalan et al. [11] reported that Wikipedia has similar accuracy and depth when compared with a professionally edited database. Reavley et al. [12] showed that the quality of information on depression and schizophrenia from Wikipedia is generally as good as, or better than, information provided by centrally controlled websites, Encyclopedia Britannica or a psychiatry textbook.

Prior work related to Wikipedia content classification aimed to automate category assignments to articles [13]. Due to diverse and fine-grained Wikipedia category labels, this task becomes very challenging. In this study, we focus on a much simpler task, which we speculate is more feasible to be integrated for improving clinical QA.

Among Wikipedia's content, its category hierarchy is widely used to produce semantic resources for different tasks, such as word disambiguation [14], topic indexing [15]. The open domain QA system Morpheus [16] integrates an ontological query model based on the Wikipedia ontology. Different from existing work, we investigated how Wikipedia category hierarchy information can be explored to identify health- and clinically-related articles from Wikipedia.

Materials and Methods

Unsupervised Approach

On Wikipedia, categories are assigned to every article, providing navigational links to all articles in a hierarchy of categories. For example, the root node "Health" is one of 22 main topic categories within Wikipedia category hierarchy and "Health Care" is one of its children. Categories and their structure are created and maintained by Wikipedia editors following the categorization guidelines. Our hypothesis is that: the more health related categories that are assigned to an article, the more likely it is that the article is health-related.

To quantify health related categories assigned to an article, we need to identify health related category subset from the Wikipedia category hierarchy. We explored a relatively simple breadth-first algorithm, which starts from the root node Health category to traverse all its children hierarchies. However, as previous study [17] has pointed out, category hierarchy in Wikipedia is not a tree structure, and some categories have multiple parent categories. Furthermore, the category structure contains cycles, which makes it problematic to extract the category hierarchy under Health category. Therefore, Health category hierarchy might contain circles such as "H3-H4-H3", and some Health related category nodes might have children nodes such as "S3" that are under another top category "Science" at a higher level, as shown in Figure 2. The former issue will lead to redundancy and even failure to traverse all health related categories; while the latter could introduce much noise by extracting non-health related categories due to network propagation. In this case, many categories in Science are not related to health.

To address the two problems, we added one rule, which extracts a category as health related only if the category is a direct or indirect child of health category and its "level" is larger than its parent node. Here we define "level" as the shortest path distance form "Health" category. Using this method we extracted 16,454 categories out of the total 742,855 descendent categories under Health. We observed that the maximum level of extracted category nodes was 11.

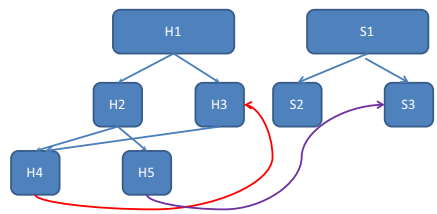

Figure 2 - Illustration of Wikipedia category network. "H" indicates Health related categories and "S" indicates Science related categories.

Based on the extracted health related category hierarchy, we developed a simple unsupervised method to identify health related articles. Specifically, we calculated health categories percentage, referred as *hcp* in (1), which measures what percentage of an article's categories are health related categories.

$$hcp = \frac{\#\ of\ \ health\ categories}{total\ \#\ of\ \ categories\ of\ \ an\ article} \quad (1)$$

Then given a Wikipedia article (w), the unsupervised approach (represented by a function f) determines whether the article is health related or not, by checking whether the *hcp* score calculated using (1) is larger than a threshold (t) or not as shown in (2) below.

$$f(w,t) = \begin{cases} 1(health) & if\ \ hcp > t \\ 0(non-health) & otherwise \end{cases} \quad (2)$$

Supervised Approach

As shown in Figure 1, we aim to perform classifications among three classes: health related clinical, health related non-clinical and non-health related. The unsupervised approach can distinguish only health-related articles from non-health related ones because no clinical related Wikipedia category hierarchy is readily available. Therefore we built supervised systems to address all the classification tasks where health category information, including *hcp*, can be integrated as addition features.

There are four classification scenarios: (a) binary classification between health and non-health articles; (b) binary classification between health related clinical and health related non-clinical articles (given that the article was health-related); (c) multiclass classification among all three classes of health related clinical, health related non-clinical and non-health; (d) pipeline system that integrate two binary classification systems to first classify an article as health or non-health and then classify the article as health related clinical or health related non-clinical if it is determined to be health related at the first step in the pipeline.

For a Bag-of-words feature, we tried using text from different sections of an article to extract unigram features, including title, title & abstract, and the full-text of the article. In Wikipedia, the abstract of the article is the text of the article before

the table of "Contents." If the article does not have a table of "Contents," the entire text of the article is considered to be the abstract. All the words were normalized by lowercasing them, removing numbers and punctuations, and stemming the words using the Porter stemmer algorithm [18]. We also removed stop words such as – 'a', 'and', 'the' etc.

For the Wikipedia category features, we speculated that categories assigned to an article represent its semantics and would aid in the classification task. Therefore we explored several features based on the health-related category hierarchy: a) health related category names; b) parent category names of health related categories; c) health category percentage; d) maximum, minimum, average and standard deviation of health related categories' depth levels in the Health category hierarchy. The hypothesis was that if an article were assigned a health related category at a deeper level in the Health category, it is more likely to be health related. We explored this hypothesis using each category feature individually and all together.

We chose Naïve Bayes Multinomial (NBM) as the supervised learning model in this study because it has shown better performance on text classification tasks (e.g. [19]). Weka machine learning toolkit [20] was used for model training and testing. For all our machine learning runs, we explored mutual information-based feature selection to further improve the classification performance and experimented with a broad range of top 10 - 11,000 unigram features. Feature section was done based on mutual information score of each feature.

Annotation and Gold Standard

To build a gold standard for training and evaluation, we developed an annotation guideline and asked each annotator to follow it and assign each article as health-related clinical, health-related non-clinical, or non-health related. For each annotation, the annotator can also enter his/her certainty level or "not sure" for uncertain cases.

We used the Wikipedia database dump files in the xml format (20120601 version accessed at http://dumps.wikimedia.org/). To validate the hypothesis in our unsupervised approach, we selected articles based on 9 different health category percentage (hcp) ranges: 0, (0, 10], (10, 20], (20, 30], (30, 40], (40, 50], (50, 60], (60, 70] and [70-100]. For each hcp range, 100 documents were randomly sampled resulting in 900 articles in total. We then recruited 9 annotators for annotation.

To minimize the amount of annotation but ensure the measurement of inter-annotator agreement, each annotator annotated a total of 120 articles with 40 articles overlapping with two other annotators (20 for each). We reported each annotator's agreement with two other annotators. Overall, 720 articles were annotated once (80 articles for each of 9 annotators) and 180 articles were annotated twice.

After the annotation, we excluded 14 (out of the 720 single annotator) articles because the certainty was marked as "not sure." For articles being annotated twice, we found 40 articles were in discrepancy, 31 of those articles were assigned consistent certainty levels (both sure or both unsure), and 9 articles were inconsistent in certainly. For a gold standard, we excluded those 31 articles with the same annotation certainty but differed in their annotations, and kept the annotations of "sure" cases when two annotators indicated different annotation certainty (one was sure and the other was unsure). The final gold standard was comprised of 855 articles, of those, 208 articles were tagged as "Health related clinical", 112 as "Health related non-clinical" and 535 as "Non-health related".

Results

Gold Standard: Agreement and Statistics

Table 1 shows the inter-agreement in our annotation using Kappa coefficient, a widely used statistical measure of inter-rater agreement taking into account the agreement by chance. We can see that using "sure" annotations improved the Kappa score across different settings, achieving the best score of 0.721 for Health/Non-health annotations.

Figure 3 shows number of articles annotated as health or non-health related as a function of incremental percentage of health-related categories assigned to articles. The results show that when the percentage increases, all health-related annotation increase, and non-health related annotation decrease.

Table 1 - Kappa Coefficient for three set of categories

Categories evaluated	Article coverage	
	All articles	*Only sure articles*
All three categories	0.595	0.649
Health/Non-health	0.664	0.721
Health related clinical/Health related non-clinical	0.535	0.615

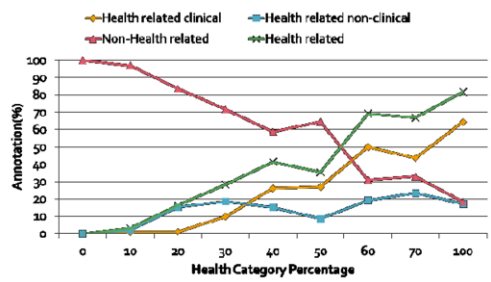

Figure 3- Percentage of annotations for different classes correlated with the percentage of health related categories assigned to the Wikipedia articles

Unsupervised baseline using health category percentage

Unsupervised baseline systems based on different thresholds were evaluated and the results are shown in Figure 4. When the threshold is 50%, i.e. the article will be tagged as health related if its health category percentage is larger than 50%, the baseline system achieved the best performance of Macro F-measure at 74.82% and Micro F-measure at 76.72%. It also validates that health category information is helpful in identifying health-related articles. Systems using thresholds less than 50% performed better than systems using thresholds larger than 50%, suggesting that articles annotated as "health related" in the gold standard typically didn't have very high health category percentages in Wikipedia.

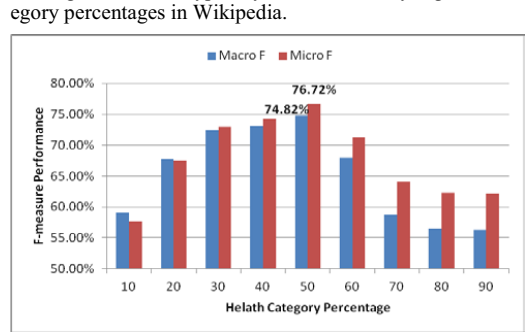

Figure 4 -Baseline system performance based on different health related percentage values

Supervised learning system

We explored three BOW feature settings: the title of the article only, title and abstract of the article, and the full-text of the article. We found that using the title and abstract of the article gave the best results. We also compared NBM classifier with Supported Vector Machines (SVM), and found that NBM slightly outperforms SVM in most settings. Due to space limitations, we only reported the results using the title and abstract of the articles with the NBM classifier. For results using feature selection, the best result is reported and values after ± indicate standard deviation in all the tables. The experimental results below are all based on 10-fold cross validation on our gold standard dataset.

Binary classification

We first attempted binary classification to classify health and non-health articles, and to classify health related clinical and health related non-clinical articles. Table 2 shows the results of classifying health and non-health articles. Here "hcp" indicates health category percentage, "cat" indicates health related category names assigned to the article, "parent" represents parent category names of the assigned health related categories, "level" means descriptive statistics of category levels in the health-related category hierarchy, and "all_cat" means using all the category features together, "FS" means feature selection, and the number of features selected in the best setting is shown in the parentheses.

Table 2- Performance of classifying health and non-health articles.

Features Used	W/O FS		With FS	
	Macro F1 (%)	*Micro F1 (%)*	*Macro F1 (%)*	*Micro F1 (%)*
BOW	85.68 ± 3.84	86.68 ± 3.56	86.04 ± 4.52	86.91 ± 4.23
BOW+ hcp	81.20 ± 2.68	81.75 ± 2.83	86.56 ± 2.98	87.32 ± 2.80
BOW + cat	85.38 ± 3.71	86.42 ± 3.43	86.17 ± 4.55	87.03 ± 4.26
BOW + parent	87.40 ± 4.55	88.24 ± 4.23	87.26 ± 4.70	88.12 ± 4.37
BOW + cat+parent	**87.91 ± 4.53**	**88.72 ± 4.23**	**87.91 ± 4.53(all)**	**88.72 ± 4.23(all)**
BOW + level	85.05 ± 4.18	86.14 ± 3.85	86.31 ± 4.64	87.14 ± 4.33
Bag of words + all_cat	85.94 ± 2.97	86.56 ± 2.96	87.89 ± 2.70	88.64 ± 2.52

We can see that for health/non-health classes, the best F1-score, 88.72%, was seen when we used bag of words with assigned category names and parent category names. This might be because the category names provided a higher level of semantics that may have helped overcome the data sparse problem. It shows that "hcp", "cat" and "level" don't seem to be helpful without feature selection, but they can make a better contribution with feature selection. This indicates that individual features in the same feature group can contribute differently. For the health related class, it achieved a precision of 86.47% and recall of 83.44%. Similar trends were found in classifying health-related clinical and health-related non-clinical articles, achieving the best results at the "BOW+cat+parent" setting resulting in a micro F1 of 89.36% and macro F1 of 88.06%. Here in Table 3 we show the best settings compared with the baseline. For the health-related clinical class, the best setting achieved the precision of 88.75% and recall of 96.62%.

We found that although each category feature type can help with performance, using all category features will introduce noise, leading to decreased performance. One difference between the two binary settings is that "hcp" and "level" are helpful with feature selection for health vs. non-health, but not for classifying health related clinical and health related non-clinical articles. We speculate that the health category percentage and level features are originated from the "Health" category in Wikipedia where no such a clinically related category hierarchy is available.

Table 3 - Performance of classifying health related clinical and health related non-clinical articles.

Features Used	W/O FS		With FS	
	Macro F1 (%)	*Micro F1 (%)*	*Macro F1 (%)*	*Micro F1 (%)*
BOW	84.21 ± 7.29	85.99 ± 6.45	85.36 ± 7.69	87.14 ± 6.45
BOW + cat+parent	88.06 ± 4.91	89.36 ± 4.40	**88.06 ± 4.91(7000)**	**89.36 ± 4.40(7000)**

Classification on all the three classes

We build two classification systems to automatically identify an article into one of the three classes: health related clinical, health related non-clinical and non-health. The first is a pipeline system which consists of two binary classifiers. The second system is a single multiclass classifier.

We observed that the single multiclass classifier achieved the best micro F1 of 85.37% and macro F1 of 78.38% among different settings, slightly outperforming the pipeline classifier's best micro F1 of 83.91% and macro F1 of 75.71%. Both achieved the best performance when selecting 500 features. In the best multi-classification setting, the precision and recall of identifying clinical health related documents is 80.56% and 93.76% respectively. Similar to previous experiments with clinical vs. non-clinical classification, "hcp" and "level" features didn't help in either multiclass classifier or pipeline classifier even with feature selection. We also found that "cat" and "parent" features provide stable performance gain, but different from previous experiments, the "BOW+parent" yielded a better performance than "BOW+cat+parent" in classifying articles into three classes.

Discussion

We explored Wikipedia's category hierarchy in our study, showing that features such as derived category names and parent names can be helpful. Note that the extraction of health related hierarchy is not error free, which might hold off the potential of related features. We also noticed that feature selection plays an important role in this task, and it chooses different numbers of features for different classification settings. Normally, binary classification would choose many more features than the multiclass classification and pipeline classification, which may be because it is prone to over fitting for binary classifications.

We conducted error analysis on misclassified articles. One type of errors was observed when the annotator incorrectly assigned the class. For example, "Patrick Mullie" is an article about an epidemiologist, who works on cancer and vitamin D. In Health vs. Non-Health classification, system identified this article correctly as "Health," while the annotator annotated it as "Non-health." Many such cases were observed when the article was about a physician or a healthcare organization. In case of health related clinical versus health related non-clinical, we also found cases where the system identified the class correctly but the annotator did not, for example "List of disorders of foot and ankle" was annotated as "Health related non-clinical" by the annotator, but the system correctly classified it as "Health related clinical."

Another type of error comes from the fact that the system relies on bag-of-word features, causing confusion for the system. For example, the article titled "total petroleum hydrocarbon" is incorrectly detected as "Health-related clinical" by the system,

because the article contains chemical text, which is common in drug names and other clinical articles.

The final limitation in our system is that it lacks the capacity of deep semantic understanding. For example, "Damping off" is a plant disease. The article on "Damping off" was incorrectly classified by the system as "Health related clinical" because the article discusses diseases, but the system cannot infer that this disease is not a human disease and so it is not related to the "Health related clinical" class based on our definition.

Conclusions and Future work

In this study, we explored textual features and Wikipedia's category hierarchy structure to identify health related content from Wikipedia. Our experiments show that features derived from the health related Wikipedia category hierarchies are helpful in identifying health related articles, but not all of them help to identify clinically relevant articles. The supervised approach that we have presented achieved promising results, suggesting that it could be effectively incorporated into existing clinical QA systems to improve the efficacy and accuracy.

For future work, we may explore additional learning features, such as syntactic parsing and Wikipedia hyperlinks, to further improve the classification performance. Furthermore, we plan to incorporate the classifier into a clinical QA system and evaluate the practical effectiveness of improving the quality of machine generated answers.

Acknowledgments

Research reported in this publication was supported in part by 1R01GM095476 to Hong Yu and by the National Center for Advancing Translational Sciences of the National Institutes of Health under award number UL1TR000161. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

[1] Moick M and Terlutter R, "Physicians' Motives for Professional Internet Use and Differences in Attitudes Toward the Internet-Informed Patient, Physician--Patient Communication, and Prescribing Behavior," *Medicine 2.0*, vol. 1, no. 2, p. e2, 2012.

[2] "Wikipedia:About - Wikipedia, the free encyclopedia." [Online]. Available: http://en.wikipedia.org/wiki/Wikipedia:About. [Accessed: 23-Jul-2012].

[3] Heilman JM, E. Kemmann, M. Bonert, A. Chatterjee, B. Ragar, G. M. Beards, D. J. Iberri, M. Harvey, B. Thomas, W. Stomp, M. F. Martone, D. J. Lodge, A. Vondracek, J. F. de Wolff, C. Liber, S. C. Grover, T. J. Vickers, B. Meskó, and M. R. Laurent, "Wikipedia: A Key Tool for Global Public Health Promotion," *Journal of Medical Internet Research*, vol. 13, no. 1, p. e14, Jan. 2011.

[4] Hughes B, I. Joshi, H. Lemonde, and J. Wareham, "Junior physician's use of Web 2.0 for information seeking and medical education: A qualitative study," *International Journal of Medical Informatics*, vol. 78, no. 10, pp. 645–655, Oct. 2009.

[5] Brokowski L and A. H. Sheehan, "Evaluation of Pharmacist Use and Perception of Wikipedia as a Drug Information Resource," *Ann Pharmacother*, vol. 43, no. 11, pp. 1912–1913, Nov. 2009.

[6] Laurent M and T. J. Vickers, "Seeking health information online: does Wikipedia matter?," *J Am Med Inform Assoc*, vol. 16, no. 4, pp. 471–479, Jul. 2009.

[7] Cao Y, F. Liu, P. Simpson, L. Antieau, A. Bennett, J. J. Cimino, J. Ely, and H. Yu, "AskHERMES: An online question answering system for complex clinical questions," *J Biomed Inform*, vol. 44, no. 2, pp. 277–288, Apr. 2011.

[8] Cairns BL, R. D. Nielsen, J. J. Masanz, J. H. Martin, M. S. Palmer, W. H. Ward, and G. K. Savova, "The MiPACQ clinical question answering system," *AMIA Annu Symp Proc*, vol. 2011, pp. 171–180, 2011.

[9] Jennifer Chu-Carroll JF, "Leveraging Wikipedia characteristics for search and candidate generation in question answering," 2011.

[10] Friedlin J and C. J. McDonald, "An evaluation of medical knowledge contained in Wikipedia and its use in the LOINC database," *J Am Med Inform Assoc*, vol. 17, no. 3, pp. 283–287, Jun. 2010.

[11] Rajagopalan MS, V. K. Khanna, Y. Leiter, M. Stott, T. N. Showalter, A. P. Dicker, and Y. R. Lawrence, "Patient-oriented cancer information on the Internet: a comparison of Wikipedia and a professionally maintained database," *J Oncol Pract*, vol. 7, no. 5, pp. 319–323, Sep. 2011.

[12] Reavley, A. J. Mackinnon, A. J. Morgan, M. Alvarez-Jimenez NJ, S. E. Hetrick, E. Killackey, B. Nelson, R. Purcell, M. B. H. Yap, and A. F. Jorm, "Quality of information sources about mental disorders: a comparison of Wikipedia with centrally controlled web and printed sources," *Psychol Med*, vol. 42, no. 8, pp. 1753–1762, Aug. 2012.

[13] Szymański J, "Towards automatic classification of Wikipedia content," in *Proceedings of the 11th international conference on Intelligent data engineering and automated learning*, Berlin, Heidelberg, 2010, pp. 102–109.

[14] Mihalcea R and A. Csomai, "Wikify!: linking documents to encyclopedic knowledge," in *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, New York, NY, USA, 2007, pp. 233–242.

[15] Medelyan O, I. H. Witten, and D. N. Milne, "Topic indexing with Wikipedia," 2008.

[16] Grant C, C. P. George, J. Gumbs, J. N. Wilson, and P. J. Dobbins, "Morpheus: a deep web question answering system," in *Proceedings of the 12th International Conference on Information Integration and Web-based Applications & Services*, 2010, pp. 841–844.

[17] Schonhofen P, "Identifying document topics using the Wikipedia category network," in *IEEE/WIC/ACM International Conference on Web Intelligence, 2006. WI 2006*, 2006, pp. 456 –462.

[18] Porter MF, "Readings in information retrieval," K. Sparck Jones and P. Willett, Eds. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1997, pp. 313–316.

[19] Agarwal S, F. Liu, and H. Yu, "Simple and efficient machine learning frameworks for identifying protein-protein interaction relevant articles and experimental methods used to study the interactions," *BMC Bioinformatics*, vol. 12, no. Suppl 8, p. S10, 2011.

[20] Hall M, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, Nov. 2009.

Address for correspondence

Feifan Liu: feifan.liu@gmail.com; Hong Yu: hongyu@uwm.edu