Domain Specific Information Retrieval and Text Mining in Medical Document

Sanghoon Lee
Department of Computer Science
Georgia State University
Atlanta, Georgia
slee172@student.gsu.edu

Yanjun Zhao
Department of Computer Science
Troy University
Troy, Alabama
yjzhao@troy.edu

Mohamed Eid Mahmoud Masoud
Department of Computer Science
Georgia State University
Atlanta, Georgia
mmasoud1@student.gsu.edu

Maria Valero
Department of Computer Science
Georgia State University
Atlanta, Georgia
maria.valero59@gmail.com

Semra Kul
Department of Computer Science
Georgia State University
Atlanta, Georgia
skul1@student.gsu.edu

Saeid Belkasim
Department of Computer Science
Georgia State University
Atlanta, Georgia
sbelkasim@cs.gsu.edu

ABSTRACT

This paper introduces a domain specific knowledge discovery technique that is applicable for both information retrieval and text mining, identifying word meanings characterized by domains. The meaning of words is identified by using a domain fusion algorithm that not only narrows domain concepts from different domain knowledge but also avoids the unknown domain problem so that specific domains can be found in a series of words. Domain knowledge is presented for the purpose of experiments on medical documents. Experiments performed over two different fields: query expansion in information retrieval and text classification in text mining, demonstrate the effectiveness of the proposed methodology.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Query formulation, Retrieval models, Search process*; I.7 [**Document and Text Processing**]: Miscellaneous

General Terms

Experimentation, Languages

Keywords

Domain, Information Retrieval, Text Mining, Query Expansion, Text Classification

BCB'15, September 09-11, 2015, Atlanta, GA, USA.
 ISBN 978-1-4503-3853-0/15/09 ...$15.00.
DOI: http://dx.doi.org/10.1145/2808719.2808726.

1. INTRODUCTION

In the fields of information retrieval and text mining, knowledge discovery from medical documents has come into the spotlight due to the unprecedented growth in both medical data volumes and biomedical literature over the last decade. A variety of methodologies for discovering knowledge from medical documents have been dedicated to not only creating innovative techniques but also making a technological breakthrough [8, 12, 22]. The methodologies are, however, mainly focused on text annotation through the use of information extraction technology using one domain knowledge, such as genes, proteins and diseases [29, 40, 44] and the experiments rely on mostly one application, for example, text categorization or text classification [21, 42, 49]. Moreover, they often do not take word meanings into consideration before analyzing text in context rather than compelling evidence of the connection between a word and a domain concept [19, 27].

Using one domain knowledge that covers a particular field of knowledge to discover a specific knowledge from medical documents may be beneficial for treating a concept representation of all the related topics. However, the concept representation is limited by a narrow range of domain knowledge. For example, medical documents relevant to health disparity may contain various topics such as particular race/ethnics, universities and regions related to health disparity, but a domain alone may not cover all of the topics due to its specialized characteristics in medical documents. This may cause unknown domain problem where knowledge leans on one side to one domain knowledge, hindering understanding of the medical documents.

In addition, the identification of word meanings can affect understanding of documents. Traditional approaches for identifying the meaning of words have been done by using definitions in a dictionary [1, 28, 37] or by applying a statistical model [7, 6, 9]. However, those approaches that rely on generalized terminologies in a dictionary are not appropriate for applying them directly on medical documents which contain specialized terminologies. This is because that the medical documents may contain complex medical terminolo-

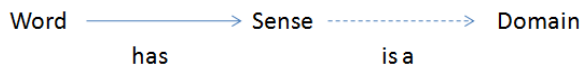

Figure 1: Word, sense, and domain representation based on ontology relations

gies as well as new medical terminologies which may not be covered by traditional dictionaries, and it can be hard to make a good achievement on a statistical model. Therefore, they still have challenges with regards to the problems that involve understanding the word meanings from medical documents.

A word often has many senses and the word senses are determined by its context. The meaning of words can be identified by determining the word senses. However, word senses described in a dictionary usually exist in a glossary form which may not suitable for applying it into applications. To address the issue of defining word senses, some studies have been done by extracting domain terminology from word senses [15, 33]. One or more glosses are determined by its context and the glosses are mapped to certain domains. Fig. 1 shows the basic idea of identifying and representing a word for a domain. The approaches have something in common with ontology. Ontology is a specification of a conceptualization and it provides a formal frame that represents a specific knowledge with a domain. We also adopt the ontology concepts so that domains are conceptualized on multi-levels.

Domain knowledge extracted by word senses can be applicable to different fields of study, such as information retrieval and text mining. Information retrieval is a well-known research area that finds useful information from large document collections with solid theoretical foundations in computer science and other sciences. Many efforts have been made to suggest various models such as Vector space model [41], Latent Semantic Indexing also known as Latent Semantic Analysis [13], Probabilistic Latent Semantic Indexing [18], and Latent Dirichlet Allocation [5] to enhance the information retrieval performance dealing with large and diverse document collections analyzing automatically. However, the representative studies have been done on pure text without any consideration of the meaning of words because they have primarily focused on creating new models to enhance retrieval performance [46, 4]. In this paper, we apply domain knowledge into the most recent topic model demonstrating that the technique with domains is adaptable to the area of information retrieval.

Text mining is a well-established research area that finds new information or high quality patterns from text by applying techniques such as, natural-language processing, machine learning and data mining. Many mining algorithms have been proposed to facilitate discovering and analyzing the patterns within large quantities of documents and combined with each other to produce a new algorithm which is more accurate or efficient than using alone [43, 20, 34, 2]. However, the algorithms also mainly focus on finding the optimum patterns by pure text with the limitation that the algorithms often ignore the meaning of words. Some researchers have introduced text mining techniques related to word senses, but their works were not to apply word meanings to algorithms but to mostly disambiguate word senses using the algorithms [48, 38]. We find word meanings with domain knowledge and apply it into a text mining technique showing the effectiveness of the use of domains.

In order to apply domain knowledge into the areas, we use two types of domains: general domain knowledge (WordNet Domains) and specific domain knowledge (Medical Subject Headings and Health Disparity Domains), used as base repositories of domain knowledge. General domain knowledge provides broad domain concepts that cover overall domain knowledge, while specific domain knowledge provides particular domain concepts that only cover special domain knowledge. We present a domain fusion algorithm that combines general domain knowledge with specific domain knowledge, providing specific domain knowledge as well as general domain knowledge for medical documents. The combination of the domain knowledge is applied to both information retrieval and text mining techniques evaluating with conventional models and algorithms respectively.

The main contributions of this paper are: 1. Our approach takes word meanings into account when discovering knowledge from medical documents. First of all, word senses are determined by its context. Second, the word senses are mapped to domains. Finally, the domain knowledge can be extracted from the medical documents. 2. We propose a domain fusion algorithm that not only narrows domain concepts from different domain knowledge but also avoids unknown domain problem. The algorithm can be applied to various fields of studies, such as information retrieval and text mining, identifying domains in a series of words. 3. We perform two experiments for demonstrating the effectiveness of the proposed methodology which is suitable for information retrieval and text mining models. The first experiment aims to determine how domain knowledge can be combined with an existing topic model in the area of information retrieval and what the expected results are. The second experiment aims to determine how domain knowledge affects on text classification methods.

The rest of the paper is organized as follows. In Section 2, we introduce two kinds of domain knowledge: general domain knowledge and specific domain knowledge. Section 3 describes domain relevance computation and domain fusion algorithm. In Section 4, we present our experiment methods and results for information retrieval and text mining. Conclusions are given in Section 5.

2. DOMAIN KNOWLEDGE INFORMATION

In this section, we introduce three domains: WordNet Domains for general domain knowledge, Medical Subject Headings (MeSH) and Health Disparity (HD) Domains for specific domain knowledge.

2.1 WordNet Domains

WordNet firstly presented by G. Miller et al. [35, 36] is a publicly available semantic lexicon of English that provides word definitions and examples of the use of the word including advantages of conventional dictionaries. A set of synonyms called Synset is used as a basic unit of WordNet and each Synset contains a brief definition called Gloss linked by semantic relations, such as hypernym, hyponym, and meronym. WordNet Domains is a lexical resource annotated by WordNet, providing semantic domain labels on word senses. WordNet Domains is structured on the basis

of 200 domains generated in a hierarchical structure semi-automatically [32]. Each sense of word is labeled with one or more domains such that domains represent senses for a particular word.

The main purpose of WordNet Domains is to provide the use of a large-scale domain application annotating with domain labels from a large domain hierarchy. In particular, it is revised by L.Bentivogli et al. [3], aiming to add some properties such as semantics, disjunction, basic coverage, and basic balancing, to WordNet Domains. Based on the Dewey Decimal Classification (DDC) system [14] which is the most widely used taxonomy for library classification system, they identified unambiguous labels avoiding label overlaps.

WordNet Domains, however, does not provide all senses for all words because it is still incomplete to link between domains senses. Also, it ignores special domains which are not specified in DDC system. In order to avoid the problems, we initially create a special definition tree that reduces gaps between domains and senses; we built HD definition tree and used it as a special domain. Next, we use two algorithms that directly link between domains and words identifying word senses.

We use WordNet Domains for our backbone domains. The main reason for using WordNet Domains is that it is applicable to wider range of tasks. Because it is built on DDC system which provides a hierarchical structure for organizing universe items, it can be considered as general domain knowledge.

2.2 Medical Subject Headings (MeSH)

MeSH is a controlled vocabulary thesaurus developed by the National Library of Medicine (NLM) [31]. It provides a hierarchical structure that covers several domains such as medicine, nursing and health care systems, consisting of headings in the twelve-level hierarchy. MeSH has been widely used for indexing biomedical articles as well as for searching medical documents. In 2014, it contains 27,149 descriptors and 218,000 entry terms indicating appropriate headings.

We utilize MeSH descriptors to cover specific domain knowledge. WordNet Domains can be used as general domain knowledge, while MeSH can be used as specific domain knowledge. Thanks to the hierarchical structure of MeSH, we adopt MeSH to represent specific domains. For example, headings such as *Cardiovascular Diseases [C14]* or *Musculoskeletal Diseases [C05]* can be the first level specific domains and specific headings such as *Heart Diseases [C14.280]* or *Bone Diseases [C05.116]* can be the second level specific domains covered by the first level specific domains. Moreover, entry terms provided by MeSH can be used for identifying specific domains in context. For example, *Cardiac Diseases* is an entry term to *Heart Diseases*.

2.3 Health Disparity (HD) Domains

Health Disparity (HD) refers to differences between groups of people with different races, ethnics and socioeconomics [10]. The differences have made severe social problems in contemporary society causing disproportionate risks for diseases. National Institute on Minority Health and Health Disparities (NIMHD) has made a lot of efforts for eliminating HD among U.S. population and has led researchers to participate in various projects related to HD producing many research documents every year. In particular, Research Portfolio Online Reporting Tools (RePORT) [1], a well-known online tool, provides researchers with efficient tool for better understanding about many National Institutes of Health (NIH) funded projects including NIMHD as well as published papers supported by NIH. In Section 4, we will discuss about the documents in more details.

Health Disparities are complex concepts that should consider many aspects such as racial, ethnic and socioeconomic status. Population groups have been considered as significant factors in HD among the aspects. We have designed HD tree based on concepts of races and ethnics. HD experts participated in our project have designed HD factors such as races, ethnics and socioeconomics and HD tree was built on the factors combining with Medical Subject Headings (MeSH) provided by NIH.

3. DOMAIN RELEVANCE AND DOMAIN FUSION

In Section 2, we have briefly introduced about three different domains which are related to general domain knowledge and specific domain knowledge respectively. In this section, we show how word senses are determined by its context and how the domains are combined with, describing domain relevance computation and domain fusion algorithm.

3.1 Domain Relevance Computation

A document may consist of sentences and a sentence may be regarded as a series of words. In order to determine word senses, their relatedness should be considered in the series of words.

Although a word sense can be represented by a domain and the meaning of words can be identified by the domains, it often contains multiple domains so that it is difficult to identify them or distinguish them without contextual clues to the meaning. Domain Relevance (DR) provides the contextual clues determining a relevance degree between domains. DR computation has been used as a critical step for better understanding of context, linking domain knowledge to words. In order to determine a word sense, we compute DR degrees generating (w, ϵ) pairs, where w indicates a word and ϵ indicates a domain, from documents.

A DR degree is determined by domain weights in a series of words. A domain weight is computed by combining two domain weights, a local domain weight and a global domain weight. A local domain weight is defined as a domain importance degree in a word and it is independent of contexts. A. Gliozzoet al. [15] presented a method that computes DR degrees deriving a domain weight from a word. We adopt the method to obtain our local domain weight. The local domain weight is computed by:

$$\omega_l = \sum_{i=1}^{n} \frac{\omega_\epsilon(i)}{N_s} \tag{1}$$

, where N_s is the number of senses in a word w and $\omega_\epsilon(i)$ is a function that represents a domain weight ω_ϵ for a sense i. n is the last sense; $\omega_\epsilon = 1/N_i$ if a domain ϵ exist in i and $\omega_\epsilon = 0$ if domains do not exist in i.

A global domain weight is defined as a domain importance degree in a window. A window indicates a length of words

[1] http://projectreporter.nih.gov/reporter.cfm

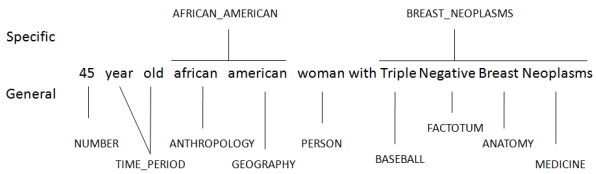

Figure 2: Example of general and specific domains

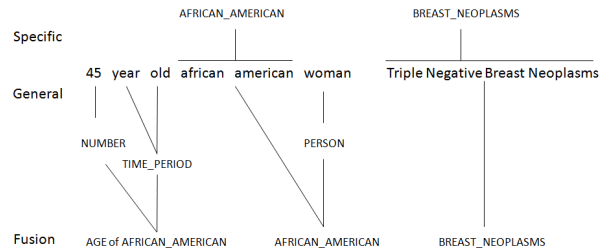

Figure 3: Example of fusion domains

in a document. It has been taken into consideration with the assumption that the narrower context is semantically more related to the target word than the broad context. A global domain assumes that a set of words in a window has more relatedness than in other window. The global domain weight is computed by:

$$\omega_g = \sum_{j=1}^{m} \frac{\omega_{\epsilon_j}}{N_w} \tag{2}$$

, where N_w is the number of words in a window w and ω_{ϵ_j} is a local domain weight in a word j and m is the last word.

3.2 Domain Fusion Algorithm

Given a series of words and domains, initial (w, ϵ) pairs can be generated by (1) and (2). However, note that (w, ϵ) pairs are wide enough to cover the meaning of words because domains are typically too broad to provide a meaningful sense. For example, a domain MEDICINE covers a large proportion of domains in medical documents but it could obstruct the findings of more specific domains, such as, adolescent medicine, behavioral medicine and emergency medicine. Furthermore, FACTOTUM, an unknown domain in WordNet Domains, could be prevailed in the (w, ϵ) pairs, impeding the ability of domain knowledge.

In order to solve the problems, we propose a Domain Fusion (DF) algorithm that not only narrows domain concepts so that specific domains can be found in a series of words but also avoids the unknown domain problem with the fusion of different domains on DR computation. DF algorithm assumes that one word has only one domain. This is because that the most appropriate word sense should represent for a word. Otherwise, inappropriate domains could decrease the accuracy of DF algorithm. The work for generating only one domain from a word can be done by (2).

DF algorithm can be used in two different domains: WordNet Domains and MeSH Domains or WordNet Domains and HD Domains. The main purpose of using DF algorithm is to narrow domain concepts from a wide or a general domain to a specific domain to which a domain priority will be given. A domain priority is a domain status that determines whether a current domain is a specific domain or not. We define WordNet Domains as a general domain and we also define HD domains as a specific domain. To help understand about the procedure of DF technique we give a serious of words as an example in Fig. 2

Fig. 2 shows (w, ϵ_g) and (w, ϵ_s) pairs: (45, NUMBER), (year, TIME-PERIOD), (old,TIME-PERIOD), (african, ANTHROPOLOGY), (american, GEOGRAPHY), (woman, SOCIOLOGY), (Triple, BASEBALL), (Negative, FACTOTUM), (Breast, ANATOMY), (Neoplasms, MEDICINE) generated by using general domains, and (african american, AFRICAN-AMERICAN), (Triple Negative Breast Neoplasms, BREAST-NEOPLASMS) generated by using specific domains. The word *with* left in Fig. 2 is a stop-word. A stop-word is a word that can be excluded from an index due to the fact that it occurs too frequently in documents and is considered as an insignificant matter for document processing. Typical stop-words are *a*, *the* and *of*. Meanwhile, the domain FACTOTUM matches to the word Negative because general domains do not have the sense for the word. Specific domains may compensate for the lack of senses. Specific domains match to their specific words but they do not have the domains for the words *45*, *year*, and *old* because, in this paper, HD Domains do not have the words. Therefore, we combine specific domains with general domains to solve the problems in both general domains and specific domains.

Fig. 3 shows (w, ϵ) pairs indicating the combination of two domains. First, the word *with* is removed because it is a stop-word and specific domains are substituted for general domains. Then, specific domains are combined with general domains left in a window. The combination should depend on a human defined rule based on a refined domain set. For example, WordNet defines the word *45* as *the cardinal number that is the sum of fourteen and one* and WordNet Domains defines it as NUMBER. However, it is not enough simply to represent the meaning of NUMBER by using words or general domains only because a machine has difficulty in identifying the meaning of NUMBER without any human defined rules. For example, the NUMBER can be identified as THEE-AGE-OF-PEOPLE, which is an element of a refined domain set, in the case of assuming that the human rule is **IF** NUMBER, TIME-PERIOD, and PERSON are in a window **THEN** NUMBER and TIME-PERIOD are THE-AGE-OFE-PERSON. Likewise, the rule can help the combination of general domains and specific domains. In Fig. 3, AFRICAN-AMERICAN followed by TIME-PERIOD can affect the meaning of NUMBER and a new domain AGE of AFRICAN-AMERICAN created by a domain user can be substituted for NUMBER TIME-PERIOD.

We define U_ϵ as a refined domain set which consists of two subsets: $U_{\epsilon_p} = \{\epsilon_1, \epsilon_2, , \epsilon_p\}$ and $U_{\epsilon_u} = \{\epsilon_1, \epsilon_2, , \epsilon_u\}$, where U_{ϵ_u} is an user-defined domain set created by a domain user manually and U_{ϵ_p} is a pre-defined domain set from existing domains. An element in U_{ϵ_u} is substituted for one or more elements in U_{ϵ_p} when it meets the rules defined by the domain user. A function: $I_P : X0, 1$, where I_P indicates whether pre-defined domains are in a window or not, is defined as:

Table 1: Domain fusion algorithm

Input: $\mathfrak{w} \in \mathcal{W}, \mathcal{L}, (w, g_\epsilon) \in G, (w, s_\epsilon) \in \mathcal{S}$
Output: $\mathcal{F}$

```
Set_g = ∅, Set_s = ∅
foreach w, w ∈ 𝔴 do
    Set_g ← (w, g_ε)
    if s_ε level ≤ ℒ
        Set_s ← (w, s_ε)
    end
end
Set_temp = ∅
Set_u = Set_g ∪ Set_s with a priority of s_ε
X ← u_ε, x ∈ X, (w, u_ε) ∈ Set_u
while I_p(x) do
    foreach X do
        Y ← X, y ∈ Y
        if I_p(y)
            Set_temp ← u_{ε_u}
        else
            Set_temp ← u_{ε_p}
        end
    end
    X = Set_temp
end
return ℱ ← X
```

$\mathcal{W}$: a set of window, $\mathcal{L}$: a domain level,
$\mathcal{G}$: a global domain set, $\mathcal{S}$: a specific domain set,
$\mathcal{F}$: a fusion domain set

$$I_p(x) \begin{cases} 1 & \text{if} \quad \forall x \quad \{x \in U_{\epsilon_p} \rightarrow x \in S_x\} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

, where x is an element of a domain set X in a window w and S_x is a subset of X. Table 1 summarizes DF algorithm.

4. EXPERIMENTS

Two experiments are performed to determine the effectiveness of the use of domain knowledge in medical documents. We perform the first experiment with Query Expansion (QE), a well-known information retrieval technique, and we perform the second experiment with Text Classification (TC), a key-technology in text mining.

4.1 Query Expansion

QE is a representative technique of information retrieval that generates alternative or expanded queries on either lexical or semantic levels for improving information retrieval performance of document collections [47, 45]. Since QE has been suggested in [24], various QE techniques have been proposed to enhance the effectiveness of information retrieval. Recently, as the volume of documents has dramatically increased in recent years, QE has received a great attention of many information retrieval communities.

In order to verify the effectiveness of the proposed domain specific knowledge discovery approach, we apply it into QE adding words with domains. The proposed approach has two advantages. First, we do not use sense definitions that yield redundant words when expanding queries. Instead, we use domain knowledge that contains refined domain concepts which avoid the problem of redundant information. Second, hypernyms and synonyms are refined by comparing with topic words which are generated by the combination of domains and a topic model called Latent Dirichelet Allocation (LDA) [5]. Due to the fact that the indiscriminate use of hypernyms and synonyms can degrade the retrieval performance, it is necessary to obtain appropriate expanded queries. Our approach for QE consists of four steps: First, find domains for words from documents. Second, generate topics from the step 1. Third, expand queries based on domains. Fourth, remove elements not relevant to topic words

First of all, we find domains for words from documents. Because the purpose of our experiment is to verify the effectiveness of domain knowledge information, we initially identify the domain knowledge from words in documents. As we described in previous section, the identification of domains is performed based on both domain relevance and domain fusion.

Second, we generate topics from documents. Topics in document collections can be represented by a set of words that shares same subject and more related to each other in the document collections. We adopt their advantage into our approach. To do this, we apply LDA, the most well-known topic model, to our approach acquiring topic words from the document collections. We use approximate inference in LDA model using the collapsed Gibbs sampling method. Gibbs sampling constructs a Markov chain computing the conditional distribution, $p(z_i z_{-i}, (w, \epsilon))$, where z_{-i} represents the topic assignments for all (w, ϵ) pairs except $(w, \epsilon)_i$. The conditional distribution is given by:

$$p(z_i = j | z_{-i}, (w, \epsilon)) \propto \frac{n_{-i,j}^{((w,\epsilon)_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \times \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,j}^{(d_i)} + K\alpha} \tag{4}$$

, where $n_{-i,j}^{(d_i)}$ is the number of (w, ϵ) assigned to topic j in document d_i excluding $(w, \epsilon)_i$. $n_{-i,}^{(d_i)}$ is the total number of (w, ϵ) in document d_i excluding $(w, \epsilon)_i$. $n_{-i,j}^{((w,\epsilon)_i)}$ is the number of (w, ϵ) assigned to topic j excluding $(w, \epsilon)_i$. $n_{-i,j}^{(\cdot)}$ is the total number of (w, ϵ) assigned to topic j excluding $(w, \epsilon)_i$. Thus, the first fraction represents the probability of $(w, \epsilon)_i$ with a topic j and the second fraction represents the probability of a topic j in a document d_i. Based on the conditional distribution given by (4), we generate top words from the result of the equation. The topic words generated by (4) will be used to remove unrelated words from the expanded query in the fourth step.

Third, we use domain knowledge information identified by the first step to expand queries. This step is different from previous approaches that expand queries by using sense definitions. Because sense definitions often contain redundant words as well as unrelated words, we use domains rather than using sense definitions. Since word senses vary in context, the identification of word sense has been considered as an important step for QE where it has a positive influence on retrieval accuracy. Our approach is used for queries as well as for document collections. Hypernyms and synonyms are generated from external resources: WordNet and MeSH. Because both WordNet and MeSH have a hierarchical structure that provides hypernyms and synonyms (entry

terms for MeSH), we can use them for QE directly. However, unrestricted use of them may cause some problems; a length of words in a query is either too long or too short to retrieve documents degrading the retrieval performance. We limit both hypernyms and synonyms to topic words generated by the second step. In the next step, we explain about it in more details.

Last, the words generated in the previous step are not always useful for retrieving documents because of the problem with the indiscriminate use of hypernyms and synonyms. It means that we need to find out a proper query by removing unnecessary words. We can remove the words less relevant to topic words by estimating $p(w|Q)$, where wis a word and Q is a query. Thanks to the theoretical foundation of information retrieval, we are able to estimate $p(w|Q)$ in document aspect using $p_d(w|Q) = \sum_{D \in C} p(w|D)p(D|Q)$, where D is a document and C is a set of documents. We define $p(w|Q)$ for topic aspect:

$$p_t(w|Q) = \sum_{T \in S} p(w|T)p(T|Q) \tag{5}$$

, where S is topics and T is topic words in S. By Bayes rule, $p(T|Q) = \frac{p(Q|T)p(T)}{p(Q)} \propto p(Q|T)p(T)$. We estimate $p_t(w|Q)$ to remove words which are less relevant to topic words generated by the second step. Thus, a query that contains both hypernyms and synonyms is refined for the use of the final query.

Our experiments are performed on OHSUMED [2] dataset that is a standard TREC collection consisting of 348,566 references which are published between 1988 and 1991. There are two reasons why we choose OHSUMED for our test collection. The first reason is that OHSUMD is widely used in benchmark evaluations of information retrieval applications. The second reason is that OHSUMED is a medical test collection in which medical terms are more informative than general terms. The dataset consists of titles and abstracts from 270 medical journals providing 63 queries with patient information. Each query was reproduced by two physicians and two medical librarians and the relevance judgments are accessed by a different group of physicians. In this paper, total 196,555 documents and 63 queries are used for the experiments. Our experiment process follows: First of all, we perform the four steps and produce new 63 queries which are expanded. Next, we compute similarities between the documents and the queries. We adopt the cosine similarity method that measures the angle between two vectors and divides the inner product of the vectors by the product of the length of vectors. The formulation is as follows:

$$sim(q,d) = \frac{q \cdot d}{|q||d|} = \frac{\sum_{k=1}^{n} q_{w_k} \times d_{w_k}}{\sqrt{\sum_{k=1}^{n} q_{w_k}^2} \times \sqrt{\sum_{k=1}^{n} d_{w_k}^2}} \tag{6}$$

, where q is an expanded query and d is a document. w is a word for the query and the document. The cosine similarity ranges from 0 to 1, meaning that it is exactly same at 1.

Last, we select 50 documents with high similarities among the documents for the performance comparison. Four different methods are compared with each other in our experiments.

- DSS-LDA: Domain Specific Search with LDA where queries are expanded by the proposed approach.
- Definition (DF) [16]: Queries are expanded by using WordNet definitions. Definitions are extracted by restricting a window and the extracted definitions are added to the original query.
- Voorhees (VO) [45]: Queries are expanded by using lexical-semantic relations. Hyponyms are added to the original query from synonyms.
- Random Indexing (RI) [26]: Queries are expanded by using RI. The closest word is added to the original query.

DSS-LDA is our model that combines Domain Specific Search with LDA. We compare it with other methods: Definitions, Voorhees and RI. Even though word sense definitions often contain redundant words, it is not surprising that the definitions are useful for information retrieval. In Guo:Semantic, they presented a semantic topic model that uses word sense definitions and showed that the word sense definitions increase the performance of topic model. We compare their method with DSS-LDA. All word sense definitions are extracted from WordNet and are used for expanding queries on the dataset. Voorhees proposed a query expansion method that utilizes semantic relations on WordNet concepts. The basic idea of the method is to add hyponyms to a query based on the semantic relations. Another method is RI that finds the meaning of words from a word space model that reduces m-dimensional word or document matrix to a new k-dimensional matrix by multiplying original matrix with a random matrix built in an incremental way. We select the method for our experiment because it is one of representative vector space techniques and can be used to find the relatedness between words statistically so that the closest word can be added to the original query.

To measure effectiveness of the methods, we use Discounted Cumulative Gain (DCG) and normalized DCG, the most popular measures of ranking quality in information retrieval [25]. DCG is used to measure the cumulative gain of the retrieved documents on their position and nDCG is used to compensate for a limitation of DCG where DCG alone cannot verify a search performance for differently sized lists of documents. DCG and nDCG are defined as follows:

$$DCG_d = count_1 + \sum_{i=1}^{d} \frac{count_i}{log_2 i} \tag{7}$$

$$nDCG_d = \frac{DCG_d}{IDCG_d} \tag{8}$$

, where d is a document rank position and $count_i$ is the number of retrieved documents in a position i. IDCG is an idealized DCG, the best result of DCG.

Fig. 4 shows the experimental results for DCG. X-axis denotes accumulated DGG and y-axis denotes the retrieved document numbers. The result shows that DSS-LDA outperforms other methods from DCG with 10 to DCG with 50. In particular, the increase rate of DCG in DSS-LDA is larger than other methods and this explains the search performance of DSS-LDA is better than others.

Fig. 5 shows the experiment result for nDCG. X-axis denotes accumulated nDCG and y-axis denotes nDCG value ranges from 0 to 1, meaning that nDCG is a perfect value

[2] http://trec.nist.gov/data/t9-filtering.html

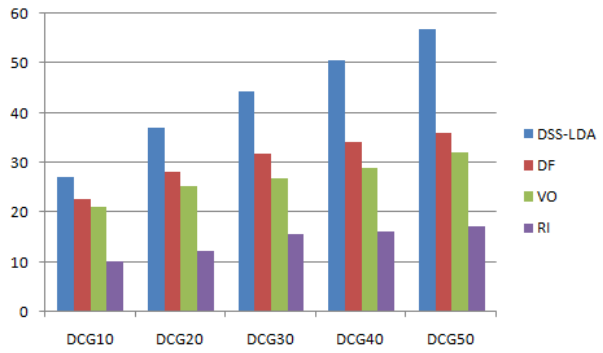

Figure 4: Experimental results for DCG

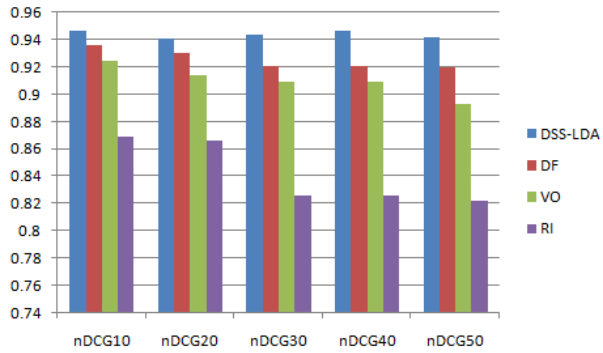

Figure 5: Experimental results for nDCG

when it is 1. The overall results show that DSS-LDA is very good in the all nDCG performance. In particular, DSS-LDA also has good results in nDCG where the number of documents is larger than 30, while others do not have. In this section, we presented a domain specific QE technique generating domain knowledge from medical documents. The experimental results showed that the proposed approach generate better results than traditional approaches. In the next section, we apply our approach to a text mining technique.

4.2 Text Classification

Text classification is a challenging and a well-studied research area that assigns documents in one or more predefined categories or classes. Existing text classification methods have been used to classify documents by subjects to facilitate a document handling process using a bag of words, given a set of labeled training documents. The difficulty with the current text classification methods is that they need a large number of labeled training documents to increase classification accuracy. Labeling training documents is very time-consuming process because it should be done by a person or an expert in the area of subjects. A bag of words causes another difficulty that a group of words share the same spelling but have different meanings. Text classification without the consideration of the meaning of words may degrade classification effectiveness or computational efficiency.

We apply DF algorithm into text classification combining WordNet Domains with HD Domains. All words in our experiment are substituted for combined domains representing word senses and the domains are used for classifying medical documents. The purpose of the experiment is to determine whether the domains without words provide better classification accuracy and performance on classification algorithms.

Four models: C4.5, NBTree, NaïveBayes and SVM, are used for evaluating the effectiveness of domains uses. C4.5 is a decision tree algorithm and our experiments were performed on J48, a Java implementation of C4.5 [39]. NaïveBayes is a well-known supervised learning algorithm that applies Bayes theorem [30]. NBTree is a hybrid version of a decision tree and naïve Bayes that generates a decision tree at the leaves [23]. Support Vector Classification (SVC) is a well-known algorithm and we use LibSVM, an open source tool [11] for our experiments. We use WEKA [17], an open source machine learning tool providing the use of the algorithms.

Two datasets of NIH project documents extracted from RePORT. The first dataset consists of six sub-datasets from National Cancer Institute (NCI), National Eye Institute (NEI), National Heart Lung and Blood Institute (NHLBI), National Human Genome Research Institute (NHGRI), National Institute of Allergy and National Infectious Diseases (NIAID) and National Mental Health (NIMH) containing two categories: with or without African American which is the third level domain in HD domains. We have collected 60 documents for each sub-dataset with a total of 360 documents in the first dataset. For each sub-dataset, 10 documents from one category are randomly extracted to build the training dataset and 20 documents are extracted for testing dataset. Likewise, 10 documents from another category are randomly extracted to build the training dataset and 20 documents are extracted for testing dataset.

In order to provide a performance assessment, our evaluation relies on two measures of performance; Accuracy and F-Measure (F1). Accuracy is a standard measure used for the binary classification performance. It depends on TP (true positive) and TN (true negative). F1 is another standard measure used to confirm classification effectiveness. It depends on TP, FP (false positive) and FN (false negative). The difference between Accuracy and F1 is that Accuracy depends on TN, while F1 does not depend on TN. It is important to take into account both measures because Accuracy can be misleading when a model with the majority negative documents achieves high classification accuracy. In that case, the model is not desirable to be used for classification. Therefore, we consider both Accuracy and F1 measure.

Table2 illustrates the performance comparison between classifiers with or without domains. According to Accuracy of the four classifiers, NBTree is the best classifier when domains are used for all documents and NBTree is the worst classifier when domains are not used for the documents. In most cases, Accuracy of the classifiers with domains is superior to the classifiers without domains, while NaiveBayes shows no significant differences between documents. The overall Accuracy of the classifiers for the documents shows that the classifiers with domains outperform the other classifiers without domains.

Fig. 6 shows the experimental results for F1. Among the results, NBTree without domains shows a slightly better result than NBTree with domains, while other algorithms with domains shows better results than the algorithms without domains. The results show that the hybrid version of two

Table 2: Accuracy for 6 groups of documents

Classifier	Domain	NCI	NEI	NHLBI	NHGRI	NIAID	NIMH
C4.5	with	**0.9**	0.6	**0.825**	**0.825**	**0.675**	**0.775**
	without	0.65	**0.725**	0.7	0.575	0.525	0.6
NBTree	with	**0.95**	**0.725**	**0.975**	**0.85**	**0.85**	**0.9**
	without	0.775	0.625	0.5	0.55	0.375	0.575
NaiveBayes	with	**0.75**	**0.675**	0.75	0.675	0.5	**0.7**
	without	0.725	**0.675**	**0.875**	**0.75**	**0.525**	**0.7**
SVM	with	**0.775**	**0.675**	**0.875**	**0.625**	**0.55**	**0.625**
	without	0.75	0.65	0.6	0.55	0.375	0.575

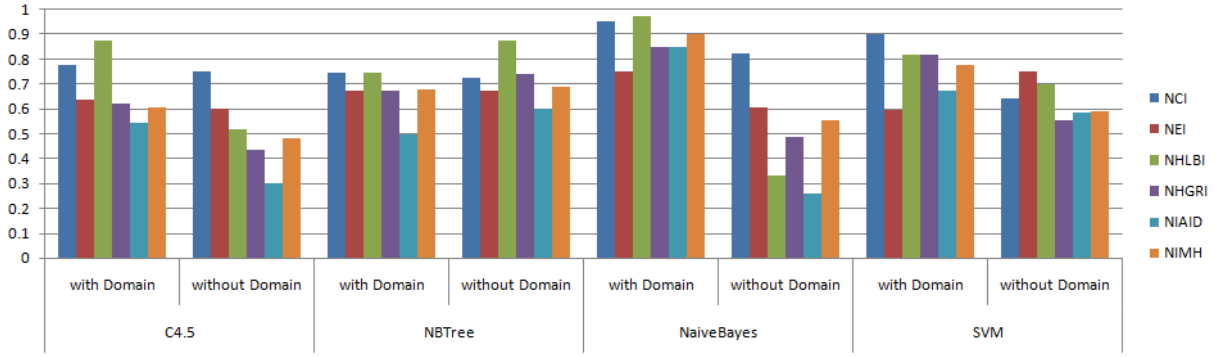

Figure 6: Experimental results for F1

Table 3: Accuracy for NIMHD

Domain	C4.5	NBTree	NaiveBayes	SVM
with	0.935	0.775	0.895	0.9
without	0.8	0.81	0.665	0.5

algorithms: C4.5 and NaiveBayes produce the opposite results compared with C4.5 or NaiveBayes. The best result on the experiment is NaiveBayes with domains in NHLBI and the worst result is NaiveBayes without domains in NIAID.

The second dataset contains two categories of African American and non African American from NIMHD. Because NIMHD is very sensitive to HD domains, it is necessary to confirm how HD domains affect documents from NIHMD. We have collected 300 documents from NIMHD projects provided by NIH RePORT and categorized them into two sets of documents; 150 documents are related to African American and 150 documents are not related to African American. For each set, 50 documents are randomly selected for a training dataset and 100 documents are selected for a testing dataset.

Table 3 illustrates the performance comparison between classifiers with or without domains. According to Accuracy of the four classifiers, C4.5 is the best classifier when domains are used for NIMHD documents and SVM is the worst classifier when domains are not used for the documents. The overall Accuracy of the classifiers shows that the classifiers with domains outperform the other classifiers without domains, while Accuracy of NBTree without domains is slightly higher than Accuracy of NBTree with domains.

Fig. 7 shows Precision, Recall, and F1 scores for NIMHD. The best F1 score is C4.5 with domains and the worst F1 score is SVM without domains. Precision, Recall, and F1 scores in NBTree without domains are slightly higher than the scores in NBTree with domains. However, the overall scores in other classifiers show that the classifiers with domains outperform the classifiers without domains.

5. CONCLUSION

We introduced a domain specific methodology for identifying the meaning words in medical documents, characterized by domains and showed that it is applicable for both an information retrieval area and a text mining area. A domain fusion algorithm is proposed not only to narrow domain concepts from different domains but also to avoid the unknown domain problem. Two experiments with the algorithm were performed over two areas: query expansion and text classification. The experimental results show that the proposed methodology produces good results on both areas.

6. REFERENCES

[1] E. Agirre, O. L. de Lacalle, and A. Soroa. Random walks for knowledge-based word sense disambiguation. *Computational Linguistics*, 40(1):57–84, 2014.

[2] L. D. Baker and A. K. McCallum. Distributional clustering of words for text classification. In *the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 96–103. ACM, 1998.

[3] L. Bentivogli, P. Forner, B. Magnini, and E. Pianta. Revising the wordnet domains hierarchy: semantics, coverage and balancing. In *the Workshop on Multilingual Linguistic Ressources*, pages 101–108. Association for Computational Linguistics, 2004.

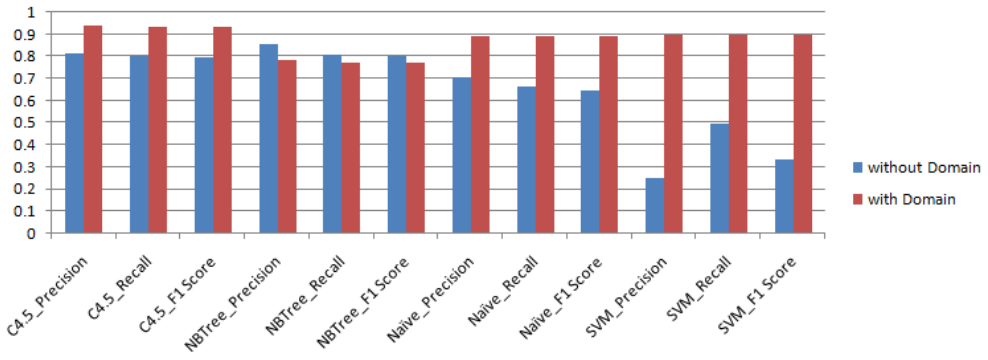

Figure 7: Experimental results for Precision, Recall, F1

[4] D. M. Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[6] J. L. Boyd-Graber, D. M. Blei, , and X. Zhu. A topic model for word sense disambiguation. In *EMNLP-CoNLL*, pages 1024–1033, 2007.

[7] S. Brody and M. Lapata. Bayesian word sense induction. In *the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 103–111. Association for Computational Linguistics, 2009.

[8] B. D. Bruijn and J. Martin. Getting to the (c) ore of knowledge: mining biomedical literature. *International journal of medical informatics*, 67(1):7–18, 2002.

[9] J. Cai, W. S. Lee, and Y. W. Teh. Improving word sense disambiguation using topic features. In *EMNLP-CoNLL*, pages 1015–1023, 2007.

[10] O. Carter-Pokras and C. Baquet. What is a.

[11] C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[12] A. M. Cohen and W. R. Hersh. A survey of current work in biomedical text mining. *Briefings in bioinformatics*, 6(1):57–71, 2005.

[13] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *JAsIs*, 41(6):391–407, 1990.

[14] M. Dewey. *Decimal Classification and Relative Index for Libraries, Clippings, Notes, Etc.* Library bureau, 1891.

[15] A. M. Gliozzo, B. Magnini, and C. Strapparava. Unsupervised domain relevance estimation for word sense disambiguation. In *EMNLP*, pages 380–387, 2004.

[16] W. Guo and M. Diab. Semantic topic models: combining word distributional statistics and dictionary definitions. In *the Conference on Empirical Methods in Natural Language Processing*, pages 552–561. Association for Computational Linguistics, 2011.

[17] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.

[18] T. Hofmann. Probabilistic latent semantic indexing. In *the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. ACM, 1999.

[19] A. Holzinger, M. G. Pinar Yildirim, and K.-M. Simonic. Quality-based knowledge discovery from medical text on the web. In *Quality Issues in the Management of Web Information*, pages 145–158. Springer Berlin Heidelberg, 2013.

[20] A. Hotho, A. Nürnberger, and G. PaaSS. A brief survey of text mining. *Ldv Forum*, 20(1):19–62, 2005.

[21] J. Huh, M. Yetisgen-Yildiz, and W. Pratt. Text classification for assisting moderators in online health communities. *Journal of biomedical informatics*, 46(6):998–1005, 2013.

[22] L. J. Jensen, J. Saric, and P. Bork. Literature mining for the biologist: from information retrieval to biological discovery. *Nature reviews genetics*, 7(2):119–129, 2006.

[23] G. H. John and P. Langley. Estimating continuous distributions in bayesian classifiers. In *Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 338–345, 1995.

[24] K. S. Jones. *Automatic keyword classification for information retrieval*. Butterworth, London, 1971.

[25] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002.

[26] P. Kanerva, J. Kristofersson, and A. Holst. Random indexing of text samples for latent semantic analysis. In *the 22nd annual conference of the cognitive science society*, volume 1036, 2000.

[27] C. S. Khoo, S. Chan, and Y. Niu. Extracting causal knowledge from a medical database using graphical patterns. In *the 38th Annual Meeting on Association for Computational Linguistics*, pages 336–343. Association for Computational Linguistics, 2000.

[28] A. Kilgarri. Senseval: An exercise in evaluating word sense disambiguation programs. In *the first international conference on language resources and evaluation*, pages 581–588, 1998.

[29] J.-D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii. Genia corpusa semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl 1):i180–i182, 2003.

[30] R. Kohavi. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Second International Conference on Knoledge Discovery and Data Mining*, pages 202–207, 1996.

[31] C. E. Lipscomb. Medical subject headings (mesh). *Bulletin of the Medical Library Association*, 88(3):265, 2000.

[32] B. Magnini and G. Cavaglia. Integrating subject field codes into wordnet. *LREC*, 2000.

[33] B. Magnini, C. Strapparava, G. Pezzulo, and A. Gliozzo. Using domain information for word sense disambiguation. In *the Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 111–114. Association for Computational Linguistics, 2001.

[34] Q. Mei and C. Zhai. Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In *the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 198–207. ACM, 2005.

[35] G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[36] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller. Introduction to wordnet: An on-line lexical database. *International journal of lexicography*, 3(4):235–244, 1990.

[37] R. Navigli. Meaningful clustering of senses helps boost word sense disambiguation performance. In *the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 105–112. Association for Computational Linguistics, 2006.

[38] R. Navigli. Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*, 41(2):10, 2009.

[39] J. R. Quinlan. *C4.5: programs for machine learning*. Elsevier, 2014.

[40] A. Roberts, R. Gaizauskas, M. Hepple, G. Demetriou, Y. Guo, I. Roberts, and A. Setzer. Building a semantically annotated corpus of clinical texts. *Journal of biomedical informatics*, 42(5):950–966, 2009.

[41] G. Salton, A. Wong, and C.-S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

[42] H. Shatkay, F. Pan, A. Rzhetsky, and W. J. Wilbur. Multi-dimensional classification of biomedical text: Toward automated, practical provision of high-utility text to diverse users. *Bioinformatics*, 24(18):2086–2093, 2008.

[43] A.-H. Tan. Text mining: The state of the art and the challenges. In *the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases*, volume 8, pages 65–70, 1999.

[44] N. Uramoto, H. Matsuzawa, T. Nagano, A. Murakami, H. Takeuchi, and K. Takeda. A text-mining system for knowledge discovery from biomedical documents. *IBM Systems Journal*, 43(3):516–533, 2004.

[45] E. M. Voorhees. Query expansion using lexical-semantic relations. In *SIGIR 94*, pages 61–69. Springer London, 1994.

[46] Q. Wang, J. Xu, H. Li, and N. Craswell. Regularized latent semantic indexing: A new approach to large-scale topic modeling. *ACM Transactions on Information Systems (TOIS)*, 31(1):5, 2013.

[47] J. Xu and W. B. Croft. Query expansion using local and global document analysis. In *the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 4–11. ACM, 1996.

[48] D. Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *the 33rd annual meeting on Association for Computational Linguistics*, pages 189–196. Association for Computational Linguistics, 1995.

[49] F. Zhu, P. Patumcharoenpol, C. Zhang, Y. Yang, J. Chan, A. Meechai, W. Vongsangnak, and B. Shen. Biomedical text mining and its applications in cancer research. *Journal of biomedical informatics*, 46(2):200–211, 2013.