Spring 5-7-2016

Multi Domain Semantic Information Retrieval Based on Topic Model

Sanghoon Lee

MULTI DOMAIN SEMANTIC INFORMATION RETRIEVAL BASED ON TOPIC MODEL

by

SANGHOON LEE

Under the Direction of Saeid Belkasim, PhD

ABSTRACT

Over the last decades, there have been remarkable shifts in the area of Information Retrieval (IR) as huge amount of information is increasingly accumulated on the Web. The gigantic information explosion increases the need for discovering new tools that retrieve meaningful knowledge from various complex information sources. Thus, techniques primarily used to search and extract important information from numerous database sources have been a key challenge in current IR systems.

Topic modeling is one of the most recent techniquesthat discover hidden thematic structures from large data collections without human supervision. Several topic models have been proposed in various fields of study and have been utilized extensively for many applications. Latent Dirichlet Allocation (LDA) is the most well-known topic model that

generates topics from large corpus of resources, such as text, images, and audio.It has been widely used in many areas in information retrieval and data mining, providing efficient way of identifying latent topics among document collections. However, LDA has a drawback that topic cohesion within a concept is attenuated when estimating infrequently occurring words. Moreover, LDAseems not to consider the meaning of words, but rather to infer hidden topics based on a statisticalapproach. However, LDA can cause either reduction in the quality of topic words or increase in loose relations between topics.

In order to solve the previous problems, we propose a domain specific topic model that combines domain concepts with LDA. Two domain specific algorithms are suggested for solving the difficulties associated with LDA. The main strength of our proposed model comes from the fact that it narrows semantic concepts from broad domain knowledge to a specific one which solves the unknown domain problem. Our proposed model is extensively tested on various applications, query expansion, classification, and summarization, to demonstrate the effectiveness of the model. Experimental results show that the proposed model significantly increasesthe performance of applications.

INDEX WORDS: Information retrieval, Semantics, Domain concepts, Topic model, Query expansion, Text classification, Text summarization

MULTI DOMAIN SEMANTIC INFORMATION RETRIEVAL BASED ON TOPIC MODEL

by

SANGHOON LEE

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

in the College of Arts and Sciences

Georgia State University

2016

MULTI DOMAIN SEMANTIC INFORMATION RETRIEVAL BASED ON TOPIC MODEL

by

SANGHOON LEE

Committee Chair: Saeid Belkasim

Committee: Raj Sunderraman

Yanqing Zhang

Hendricus Van Der Holst

Electronic Version Approved:

Office of Graduate Studies

College of Arts and Sciences

Georgia State University

May 2016

DEDICATION

I dedicate this dissertation to my wife Mijeong Oh for her ongoing love and support.

ACKNOWLEDGEMENTS

I would like to gratefully and sincerely thank my advisor, Dr. Saeid Belkasim, for his valuable guidance and suggestions throughout my research work. His clear perspective and great support inspired me to carry the torch of knowledge and love of learning.

I would also like to thank my committee members, Dr. Rajshekhar Sunderraman, Dr. Yanqing Zhang, and Dr. Hendricus Van der Holst. They always encouraged me to do my best in the dissertation work.

I would especially like to thank my colleagues, Yanjun Zhao, Semra Kul, Mohamed Masoud, Maria Valero, Stacey Levine, Janani Balaji, Sunny Shakya, Satish Puri, Sanish Rai, Zhiyi Wang, Guoliang Liu, Peisheng Wu, Mingyuan Yan, Long Ma, Yunmei Lu, and Dhara Shah for the useful discussions related to the research work. Thank you for all of the meetings and chats over the years.

Also, I would like to thank my RTEMD members, Dr. Seung-Jin Moon, Chan il Park, Younghun Chae, and Jihoon Yun for all their help and guidance. All the support they have provided me over the years was the greatest gift.

Finally, I would like to thank my wife, Mijeong Oh, for all of the sacrifices that she has made on my behalf. I can't thank her enough for her love and support throughout my life. Words cannot express how grateful I am to her family as well as to my family, and especially to my mother who couldn't see this dissertation completed. Thank you for supporting me for everything.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

1.1 Background and motivations

Information Retrieval (IR) is a well-established research area in computer science.The idea of IR is credited to Vannevar Bush after publishing his essay“As We May Think” in 1945. Bush introduced a concept of IR system called as Memex that enables individuals to read and write content on a large scaled data. He described that Memex would operate as an indexed repository of knowledge and carry out a sequence of work faster than human experts. This essay has significant influence on contemporary researchers seeking relevant information from various resources such as text, audio, and images [1-2]. Since then, a great deal of effort to improve IR strategies has been exerted by many researchers.

IR strategies have been established in five well-known theoretical models: Vector Space Model (VSM), Latent Semantic Analysis (LSA), Random Indexing (RI), Probabilistic Latent Semantic Indexing (PLSI), and Latent Dirichlet Allocation (LDA). Among the models, LDA has recently received most attention in many research communities because of its advantages that enable readers to advance their understanding as well as discover of hidden topics from large document collections. LDAhas been widely used as a topic model in text document analysis to generate topic words from text corpora with statistical relationships between words in context.

Topic models provide an efficient processing of text corpora, but they fall apart when words occur infrequently in document collections. This failure is due to the fact that these modelsinfer hidden topics from documents based on statistics rather than understanding word meanings. Moreover, IR performance is often degraded when using these models directly without any consideration of the meaning of words [7,8].

Identifying the meaning of words in context is not difficult for human interpreters, but remains a challenge for even the most advanced machinessince a word has multiple senses indicating different meanings in different contexts. A domain can be defined as a particular field of knowledge that represents a particularconcept of all related topics. Using domains for identifying the meaning of words in context can be a solution for this challenge, providing a structural view of specific word spaces [9, 10].

In this dissertation, we propose a new domain specific topic model that combines domain concepts with a topic model, identifying word senses as well as generating topic words from text document. The proposed model provides two domain specific algorithms: domain relevance algorithm and domain fusion algorithm. The algorithms not only narrow domain concepts from broad domain knowledge but also attenuate an unknown domain problem.

This dissertation is organized as follows: Chapter 2 describes theoretical background of Information Retrieval (IR) and Chapter 3 explains research works closely related to our model. In Chapter 4, we present the proposed novel domain specific topic model. Chapter 5 introduces a new medical document retrieval application based onour proposed domain specific topic model. Chapter 6 presents a new text summarizationmethod as an application to the domain specific topic model. Chapter 7 explains a new tag based image retrieval methodalso as an application of the domain specific topic model. Chapter 8 provides the conclusion and future works.

2 THEORETICAL BACKGROUND

2.1 Vector space model

Vector Space Model (VSM) is a widely used IR model that represents a query and a document as a set of vectors. VSM was introduced by Salton [3] to be used for the Mechanical Analysis and Retrieval of Text (SMART) information retrieval system. SMART system has

great influence on today's search engines including many fundamentalconcepts such as VSM, Relevance Feedback (RF), and Rocchio classification.

The basic premise of VSM on querying documents is that if q is closer to d_1 than d_2, then the query is more relevant to d_1, where q is a query vector, d_1 is the first document vector, and d_2 is the second document vector.

Theyare formally defined by:

$$\mathrm{q} = \left(\mathrm{t}_{1,\mathrm{q}}, \mathrm{t}_{2,\mathrm{q}}, \mathrm{t}_{3,\mathrm{q}}, \ldots, \mathrm{t}_{\mathrm{n},\mathrm{q}}\right) \tag{2.1}$$

$$\mathrm{d}_{\mathrm{i}} = \left(\mathrm{t}_{1,\mathrm{i}}, \mathrm{t}_{2,\mathrm{i}}, \mathrm{t}_{3,\mathrm{i}}, \ldots, \mathrm{t}_{\mathrm{m},\mathrm{i}}\right)$$

, where q is a query vector, d_i is a i-*th* document vector, and t is a weight for unique term.n and m are the number of unique terms in q and d_i respectively.

To compare the relevance between a query and a document the cosine similaritycan becomputedby:

$$\mathrm{sim}(\mathrm{q}, \mathrm{d}) = \frac{\mathrm{q} \cdot \mathrm{d}}{|\mathrm{q}||\mathrm{d}|} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{k}} \mathrm{t}_{\mathrm{i},\mathrm{d}} \times \mathrm{t}_{\mathrm{i},\mathrm{q}}}{\sqrt{\sum_{\mathrm{i}=1}^{\mathrm{k}} \mathrm{t}_{\mathrm{i},\mathrm{d}}^2} \times \sqrt{\sum_{\mathrm{i}=1}^{\mathrm{k}} \mathrm{t}_{\mathrm{i},\mathrm{q}}^2}} \tag{2.2}$$

, where $\mathrm{q} \cdot \mathrm{d}$ is a dot product of the query vector and the document vector. $|\mathrm{q}|$is a norm of the query vector and $|\mathrm{d}|$ is a norm of the document vector. k is the number of unique terms.$\mathrm{sim}(\mathrm{q}, \mathrm{d}) = 1$ when q is equal to d and $\mathrm{sim}(\mathrm{q}, \mathrm{d}) = 0$ when q has no terms ond.

VSM has gained in popularity because of the convenience of computingsimilaritiesbetween documents.However, VSM often takes a lot of time to compute a high dimensional spacein which a huge amount of different termsexist. Moreover, VSM ignores semantic relationships between terms and does not preserve any sequential order in a given document.

2.2 Latent semantic analysis

Latent Semantic Analysis (LSA) [4] is a well-known statistical model that analyzes co-occurrence of terms in a set of documents. The basic idea of LSA is that terms that co-occur frequently in similar contexts are more semantically related than others. LSA includes Singular Value Decomposition (SVD), a dimension reduction technique thattransforms a standard co-occurrence matrix into a much smaller and denser representation. Terms and documents are corresponded to rows and columns of the matrix. SVD satisfies the following relation:

$$M = U\Sigma V^{*} \tag{2.3}$$

, where U and V are orthogonal matrices for a matrix MwhileΣis the diagonal matrix that contains singular values of M. Low-dimensional latent vectors can be obtained by computing meaningful association values between documents when the lower values of Σ are removed from the original values ofΣ. This means that terms that appear in a document can be represented as meaningful terms of another document that does not have the same terms.

LSA has been widely used in many information retrieval applications [33-35] because of its several attractive properties.LSA locatesboth documents and terms in a same concept space so that it is possible to compute a distance between two semantically related documents. Thus, LSAhas been used for many applications such as clustering, classification, and cross-language IR, facilitating the use of concept space.However, LSA requires large memory space because ofa characteristic of SVD which useswholespacewhen analyzinga set of documents.Moreover, it is often very difficult to determine an optimal dimension size to performSVD.

2.3 Random indexing

Random Indexing (RI) [11] is a distributional statistic model that extracts similar terms from a set of documents based on sparse distributed term representations. RI is a scalable alternative to LSA, which avoids computational cost of a matrix factorization. The basic idea of RI is that a high dimensional model is randomly projected into a low dimension one.

RIaccumulates context vectors with the assumption that terms that occur in a same context tend to have similar meanings. RI reduces anm-dimensional matrix to a new k-dimensional matrix by multiplying an original matrix randomly in an incremental way. The model satisfies the following relation:

$$F_{n\times m}R_{m\times k} = F'_{n\times k} \tag{2.4}$$

, where F is a given matrix and R is a random matrix. RI has two-step operations. First, high-dimensional random vectors (index vectors) are assigned to each context, consisting of randomly distributed small numbers (+1 and -1, and 0). This means that values are distributed in a random way but the number of two values (+1 and -1) is smaller. Next, the vector space representation of a term or a document is obtained by summing the context vector for the term or the document. This means that the context vectors can be utilized for similarity computation even if there are small examples encountered.

RI provides a scalable dimension reduction technique to avoid the computation of whole space in a set of documents.Thus, RI does not require a significant computational power in IR.However, optimal parameters in RI should be predetermined to be used in many applications, and still requires an intensive processing power when a large number of documents are involved in the model.

2.4 Probabilistic latent semantic analysis

Probabilistic Latent Semantic Analysis (PLSA) [5] is oneof topic models that discover topic structures behind words froma set of document. The basic idea of PLSA is associating unobserved variables (topics) with other observable variables (terms and documents).

A joint probability of an observed pair $P(d, t)$ is defined by:

$$P(d, t) = P(d)P(t|d)$$

$$P(t|d) = \sum_{z \in Z} P(t|z)P(z|d) \quad (2.5)$$

,where t is a term and d is a document assuming that t and d are conditionally independent given a latent variable z. Then, the model is parameterized by:

$$P(d, t) = \sum_{z \in Z} P(z)P(d|z)P(t|z) \quad (2.6)$$

To estimate the latent variable models, it uses the Expectation Maximization (EM) algorithm [30]. In an expectation step, posterior probabilities are computed for the latent variables. In a maximization step, parameters are updated.

Probabilistic retrieval techniques have been widely used in improvingIR systems since it isconveniently applied to various models [31, 32]. PLSA outperforms LDA, generating hidden topics maximizing its predictive power. However, PLSA requires many parameters depending on the number of documents, causing overfitting problem.

2.5 Latent dirichelet allocation

Latent Dirichlet Allocation (LDA) [6] is currently the most common topic model that generates specific topics from a set of documents. The basic idea of LDA is that documentsare modeled as a mixture of multipletopics and each topic is represented as a distribution over the words. A generative process of LDA is as follows:

First, a sequence of Nwords is drawn from Poisson Distribution.

Second, a k-dimensional random variable θ is drawn from a Dirichlet prior with α.

Third, for each of the N words w_n:

- A topic $z_n$is drawn from Multinomial(θ).
- A word $w_n$is drawn from $p(w_n|z_n, \beta)$, a multinomial probability conditioned on the topic z_n and β.

The Poisson distribution is a discrete probability distribution that represents a probability of events occurring in a certain period of time or space. LDA assumes that a document is a sequence of N words drawn from the Poisson distribution generating document length distributions from a corpus. αis a K-vector showing how much a Dirichlet prior scatters around different topics. A Dirichletdistribution is a probability distribution over the space of multinomial distributions [29]. Because the Dirichletdistribution is conjugate to the multinomial distribution, it can be conveniently used to compute the posterior distribution. Bis a K × Vmatrix, where $\beta_{ki} = p(w^i|z^k)$ and Vis the total indexed vocabulary.

The LDA seeks the model parameters α and β that maximize the likelihood $p(D|\alpha, \beta) = \prod_{d=1}^{M} p(w_d|\alpha, \beta)$. The key part of LDA is to compute the posterior distribution $p(\theta, z|w, \alpha, \beta)$finding the distributions of hidden variables θ and z.

However, the distributionsof θ and zare intractable because a coupling between θ and β in the summation over latent topics occurs when computing $p(w|\alpha, \beta)$. This is whyLDA uses an approximate inference algorithm that maximizes likelihood of the lower bound. LDA uses the variational EM procedure to maximize a lower bound about the variational parameters γand φ and it maximizes a lower bound about the parameters αand β, and then, LDA finally finds the distributionθ and z.

LDA can be viewed as a modification of PLSA, allowing us to use apriori information about document collections, narrowing down the list of topics by additional control parameters. Unlike PLSA, LDA alleviates overfitting problems using the variational inference approach and provides a powerful module that is adapted in many complicated models. However, LDA still remains a problem that topic words generated from LDA arenot related to have cohesion within topics semantically strong enough.

2.6 Summary

In this chapter we described five differentIR models: VSM, LSA, RI, PLSA, and LDA,to introduce the theoretical backgrounds of the proposed model. Among them, LDA has come into the spotlight due to its adaptable characteristicof generating hidden topics from unstructured documents.

3 RELATED WORKS

In this chapter, we describe recent researches closely related to our dissertation. Since topic models have been combined with various research topics, it is very difficult to describe them in a single research stream. Thus, we categorize them into three groups: word sense disambiguation with topic models, semantics on topic models, and topic models of language processing application.

3.1 Word sense disambiguation with topic models

Word Sense Disambiguation (WSD) is a very challenging technique that disambiguates word senses in a given context. Unlike humans that determine the meaning of words in context without much difficulty, machines may encounter a problem in identifying the meaning of words because words often have more than one meaning. Many efforts have been made to tackle this problem using topic models [21-24, 36, 37].

Boyd-Graber and Blei [36] proposed an unsupervised approach that combines a topic model with a WSD technique [38] to find predominant senses for nouns. They used word senses as additionallatent variables on a topic model and showed that their approach improves the performance of WDS. Cai et al. [23] introducedasupervised approach that exploits topic features for disambiguating word senses in context. They trained a topic model from unlabelled data to generate the topic features, and then the generated topic features are used in a supervised system. They showed that context information derived from a topic model significantly improve WSD accuracy. Brody and Lapata [21] presented a word sense induction technique that models contexts as samples from a multinomial distribution space over senses. They maintained that contexts surrounding ambiguous wordsnot only reflectthe meaning of wordsbut also generate meaningful words from local topics.Yao and Durme [39] proposed a nonparametric Bayesian model that induces senses from words automatically. The basic idea of the model is that the number of sense clusters is automatically decided to avoid a limitation on fixing the number of senses. Assuming that word senses are determined by its contextual information, they showed that the proposed model leads to similar results compared with the model of Brody and Lapata [21].

WSD associated with topic models have aimed to find possible word senses in context by integrating different linguistic features. However, these approaches merely usetopic models for WSDfocusing on disambiguating word senses and do not enhance the performance of topic models.

3.2 Semantics on topic models

A word “semantic” has various meanings in different area [41-46]. In this dissertation, we use Lyons’s definition: “Semantics is the study of meaning” [41], and restrict the study to words. Thus, we define “semantic” as “the study of word meanings in context”.

Several studies on semantics have been done by using topic models. Chemudugunta et al. [25] proposed a probabilistic framework that combines semantic concepts with a statistical topic model. They built some semantic conceptsin a form of ontologicalconcepts derived from human-defined concepts, and then used the concepts to derive topics assigned to words. They extended the framework by combining topics with hierarchical concepts [40], and showed that hierarchal concepts improve a quality of topic models. However, their works remain unclear because it needs to explain how to define human knowledge as well as relations between concepts. Moreover, they do not show any applicable task for the framework. Recently, WeiweiGuo and Mona Diab [26] presented a semantic topic model that uses definitions of a dictionary. They modified LDA to create a new semantic topic model, and showed that their modelimproves classification accuracy. However,their model has a drawback in terms of that dictionary definitions are too sensitive to accomplish different types of tasks with different dictionaries.

3.3 Topic models of language processing application

Topic models have been used in a variety of language processing applications [16, 20, 27, 28, 47-51]. Wei and Croft [50] proposed a LDA based document retrieval model that applies LDA into an ad-hoc retrieval application. They combined a document model that estimates the maximum likelihood of a word in a document with LDA, constructing the LDA-based document retrieval model. They showed that the LDA-based documentretrieval model outperforms acluster-based retrieval model in ad-hoc retrieval application. D. Andrzejewski and D. Buttler

[51] presented a relevance feedback technique that uses latent topics as users' feedbacks. They allowed users to provide their feedbacks at the latent topic level of LDA. Their experimental results showed that the usage of topicswith feedbacks improve IR performance. Their work remained another potential IRmechanism called a query expansion that generates alternative queries for users. The query expansion techniques with topic modelshave providedbetter results onIR. WangandTanaka[27] presented a query expansion technique that generates queries from clustering results. However, their strategy only takes word similarities into consideration for obtaining clusters withoutidentification of word senses. Zeng QT et al. [28] proposed three differentquery expansion methods in the area of clinical research. They used synonyms, a trained topic model, and related words for expanding queries.They determined that the query expansion witha topic model producesthe best results among them.

3.4 Summary

Many research works closely connected to topic models have been proposed for improving IR performance. Several techniques have contributed to considerable improvements in the areas of WSD, semantics, and language processing applications. In the next chapter, we will presentour proposed model.

4 DOMAIN SPECIFIC TOPIC MODEL

In this chapter, we propose a new domain specific topic model that combines domain concepts with a topic model. In Section 4.1, weintroduce domain concepts. In Section 4.2 and 4.3, we present two novel domain specific algorithms: domain relevance algorithm and domain fusion algorithm. In Section 4.4, we describe a combination of domain concepts and a topic model. Section 4.5 summaries this chapter.

4.1 WordNet and WordNet Domains

WordNet is a publicly available semantic lexicon that includes definitions of words and word usage examples [12]. While WordNet is very similar to conventional dictionaries, there are some distinct differences: specific word senses are identified by interconnecting synsets, basic units of WordNet, and semantic relations between words are provided in WordNet. Total 117,000 synsets are linked to each other with conceptual relations, such as IS-A, HYPERNYM, and HYPONYM.

WordNet has been widely used in manyresearch works because ofits generalized knowledge base that can be applied for any domain. WordNet Domains[1] is a structured lexical resourcethat provides semantic domain labels providing the generality of WordNet. As part of "The WordNet Domains Project" which links WordNet to domains, WordNet Domains was developed to provide use of large-scale domain applications with domain labels. Particularly, L.Bentivogli et al. [52] added several properties: semantics, disjunction, basic coverage, and basic balancing to WordNet Domains. Dewey Decimal Classification (DDC) system [53], the most widely used taxonomy for library classification system,was involved to identify unambiguous labels avoiding label overlaps.

Table 4.1WordNet Domains for a word "black"

Sense	Synset and Gloss	Domains
#1	black, blackness -- (the quality or state of the achromatic color of least lightness (bearing the least resemblance to white))	COLOR
#2	total darkness, lightlessness, blackness, pitch blackness, black -- (total absence of light; "they fumbled around in total darkness"; "in the black of night")	FACTOTUM
#3	Black, Joseph Black -- (British chemist who identified carbon dioxide and who formulated the concepts of specific heat and latent heat (1728-1799))	CHEMISTRY
#4	Black, Shirley Temple Black, Shirley Temple -- (popular child	THEATRE

[1] http://wndomains.fbk.eu

	actress of the 1930's (born 1927))	
#5	Black, Black person, Negro, Negroid -- (a person with dark skin who comes from Africa (or whose ancestors came from Africa))	ANTHROPOLOGY
#6	black -- ((board games) the darker pieces)	CHESS
#7	black -- (black clothing (worn as a sign of mourning); "the widow wore black")	FASHION, RELIGION

WordNet Domains is structuredon the basis of 200 domains generated in a hierarchical structure [6]. Each sense of a word is labeled with one or more domains and FACTOTUM, a domain name, is used for a special case of domain that is unknown. Table 4.1 shows word senses and labeled domains for a word “black”.

In this dissertation, we define a domain as a particular field of knowledge that represents concepts of all related topics. Generally, a domain has various notions. For example, a domaincan bean area of interest or a particular person or organization. Moreover, a domain can be a set of possible quantities. However, these notions are often ambiguous when identifying domain concepts in a specific use of domains. Therefore, we use our domain definition throughout the dissertation.

We use WordNet Domains for our general domain. The main reason for using WordNet Domains is that WordNet Domains can be applicable to wide range of applications since it follows the generality of WordNet by labeling domains. Moreover, WordNet Domains is built on DDS system which provides a hierarchical structure for organizing universe items, thus we can cover general domain concepts.

4.2 Domain relevance algorithm

Domain Relevance (DR) is a key measure of determininga degree of relatedness between domains. As domain-relatednessin context affects a predictable concept of domains, we

cangenerate this concept of domainsby computing DR. However, DR without consideration of word senses may not represent domain concepts correctly because words are often associated with many senses related to different domains, thus producing improper domain-relatedness degrees. We propose a DR algorithm that finds (word, domain) pairs in which domains have the highest weights for each word by computing domain-relatedness given a series of words. DR algorithm generates(w, ε) pairs, where w is a word and ε is a domain. These pairs are used as initial(w, ε) pairs on Domain Fusion algorithm that will be explained in Section 4.3.

A domain weight is a measure of indicating how much a domain has in common in context. We compute a domainweightto find a domain-relatedness. Two domain weights, a local domain weight and a global domain weight, are combined to generate a domain weight for a domain.

The local domain weight is used to emphasize the importance of word domain independently on contexts. A. Gliozzoet et al. [9] presented a domain relevance estimation method that derives a domain weight from a word. We adopt their method to obtain our local domain weight. The local domain weight is computed by:

$$\mathrm{L}(\varepsilon_{\mathrm{k}}) = \frac{\sum_{i=1}^{\mathrm{N_s}} \mathrm{f}(i)}{\mathrm{N_s}}, \mathrm{k} = 1,2, \dots, \mathrm{n} \qquad (4.1)$$

, whereε is a domain without overlap between domains. n is the number of domains and$\mathrm{N_s}$ is the total number of senses in a word.f(i) is a function that represents a domain weight ω_ε for a sense i; f(i) $= 1/\mathrm{N}_{\varepsilon_i}$if a domain ε exist in$i$and $\omega_\varepsilon = 0$ if domains do not exist in$i$.$\mathrm{N}_{\varepsilon_i}$ is the number of domains in i.

We define a global domain weight as a measure of the importance of a domain in a window. A window represents words in a range of document.It is very important that the length of a document can affect domain weights in the sense that domains in narrower context

aresemantically more related than domains in broad context. We assume that words in a window have more relatedness than others in another window. A global domain weight is computed by:

$$G(\varepsilon_k) = \frac{\sum_{j=1}^{N_w} L(\varepsilon_k)_j}{N_w}, k = 1,2, \dots, m \quad (4.2)$$

, whereε is a domain without overlap between domains. m is the number of domains and N_w is the total number of words in a window.

Given a set of words and a set of domains, our DR algorithm finds(w, ε) pairs. Algorithm 4.1 summarizes our DR algorithm.

Algorithm 4.1: Domain Relevance

Input: $w \in S_w$, $\varepsilon \in S_\varepsilon$

Output: $S_{DR} = \{(w_1, \varepsilon_i), (w_2, \varepsilon_j), \dots, (w_n, \varepsilon_k)\}$

1. while ε in w
2. $S_{Local} \leftarrow (w, \varepsilon, L(\varepsilon))$
3. end
4. while S_{Local}
5. $S_{Global} \leftarrow (w, \varepsilon, G(\varepsilon))$
6. end
7. $S_{DR} \leftarrow (w, \varepsilon)$with the highest $G(\varepsilon)$ for w in S_{Global}
8. return S_{DR}

S_w is a set of words, $S_\varepsilon$is a set of domains, and $S_{DR}$is a set of (w, ε) pairs. S_{Local} is a set of three pairs: a word, a domain, and a local domain weight. S_{Global} is a set of three pairs: a word, a domain, and a global domain weight.

4.3 Domain fusion algorithm

Domains in WordNet Domainsare general enough to include a broad field of knowledge, but they are not appropriate to be used for a specific range of knowledge. For example, a domain "MEDICINE" covers a large proportion of domains in medical documents, butit may not be used for identifying specific domain knowledge, such as Heart Disease, Medical Device, and Health Disparity. Furthermore, an unknown domain, for example "FACTOTUM" defined by WordNet Domains,can be prevailed in documents when words have no domains. This may decreasethe quality of word sense identification. In order to avoidthese problems, we propose a Domain Fusion (DF) algorithm that not only narrows domain concepts but also avoids unknown domain problem.

DF algorithm takes (w, ε) pairs generated by Algorithm 4.1. Each pair of (w, ε) indicates that a word matches a domain which is the most weighted among other domains.We assume that a word sense can be represented by the most weighted domain in a word.In order to narrow domain concepts in a document, DF algorithm borrows a priority concept that gives the right to one before others. We give a special priority for a specific field of knowledge to narrow domain concepts. Because general domains representing a general field of knowledge cannot cover the specific field of knowledge, it is necessary to narrow the concept of domains by adding specific domains.

A specific field of knowledge can be illustrated by an example from Medical Subject Headings (MeSH). MeSH is a well-known vocabulary thesaurus provided by National Library of Medicine (NLM), and has been used for searching the scientific literature of medicine. Sixteen main branches: Anatomy, Organisms, Diseases, Chemical and Drugs, Analytical and etc. of MeSH tree contain their sub branches specifying their specific field of knowledge. For example,

Cardiovascular Diseases is a child tree of Diseases and includes Cardiovascular Abnormalities, Cardiovascular Infections, Heart Diseases and etc. Thus, the range of a specific knowledge field is determined in MeSH.

We introduce DF algorithm by describing how two different domain knowledge: WordNet Domains and MeSH, are combined with each other. As we described in Section 4.1, WordNet Domains providesgeneral domains to cover most of common domains in documents. However, it cannot be used for a specific field of knowledge. MeSH can specify a particular field of knowledge related to medical documents, but it does not include broad domains such as ART, HISTORY, and SPORT. WordNet Domains is used as general domains and MeSH[2]is used as specific domains.

We have generated(w, ε) pairs from a document by using Algorithm 4.1 and have chosen WordNet Domains and MeSH as two different domain knowledge in DF algorithm. Now we describe DF algorithm with concrete cases.

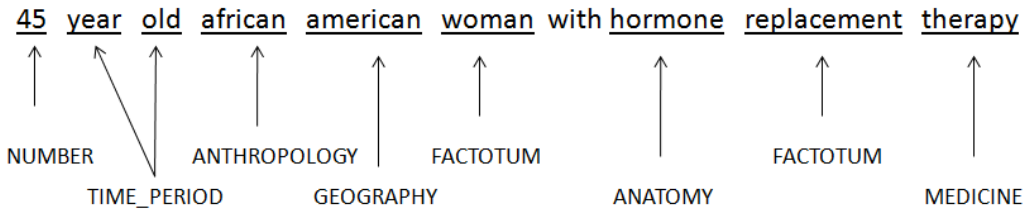

Figure 4.1Word-domain pairsgenerated by Algorithm 5.1 with general domain knowledge

Figure 4.1 shows an example of(w, ε_g) pairs: (45, NUMBER), (year, TIME_PERIOD), (old, TIME_PERIOD), (african, ANTHROPOLOGY), (american, GEOGRAPHY), (woman, FACTOTUM), (hormone, ANATOMY), (replacement, FACTOTUM), (therapy, MEDICINE)generated by Algorithm 4.1 with general domain knowledge. Each word matches

[2]http://www.nlm.nih.gov/mesh

one domain if a word contains at least one domain, excepting a word in a stop-word list. A word "with" does not match a domain because it is a stop-word. A stop-word is a word that is filtered out from a document because it occurs too frequently in a document and it does not have meaningful information. A lot of studies have maintained that the removal of stop-words from a document improves IR performance. Therefore, DF algorithm is performed by removing a list of stop-words. A typical stop-word list includes words such as "a", "the", "of", and so on. Meanwhile, due to the fact that general domains contain an unknown domain "FACTOTUM", it is not unusual for "woman" and "replacement" to be matched to "FACTOTUM" which is not dealt with in any meaningful way. This is another problem of general domains that we already discussed about it at the beginning of this section. DF algorithm avoids the problem combining specific domains with general domains.

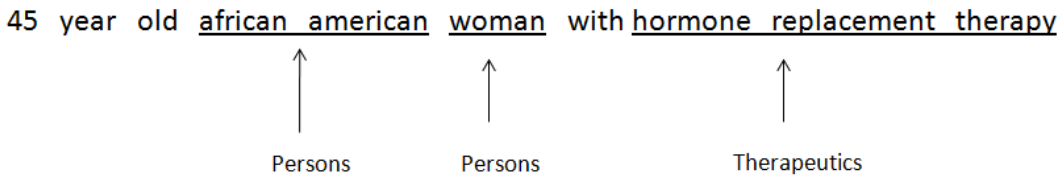

Figure 4.2Word-domain pairs generated by Algorithm 5.1 with special domain knowledge

Figure 4.2 shows an example of (w, ε_s) pairs: (africanamerican, Persons), (woman, Persons), (hormonereplacementtherapy, Therapeutics)generated by Algorithm 4.1 with specific domain knowledge.Each wordmatches one domain if a word contains at least one domain, excepting a word in a stop-word list in the same manner as Figure 4.1. We use children of sixteen main branches of MeSH as an example of specific domains. For example, "Persons (M01)" is a child of "Named Group (M)" and "Therapeutics (E02)" is a child of "Analytical, Diagnostic and Therapeutic Techniques and Equipment (E)". Each specific domain matches their words, but "45", "year", and "old" do not have their domain because the words are not defined in

MeSH. Thus, it is necessary to compensate and allow for the words by assigning general domains. We combine general domains with specific domains by givinga priority to specific domains.

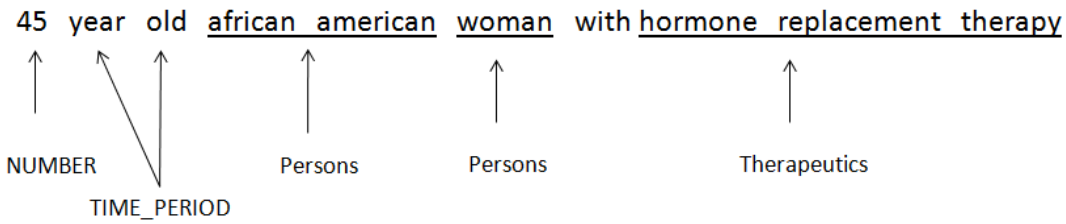

Figure 4.3 Word-domain pairs generated by combining general domains with specific domains

Figure 4.3 shows an example of(w, ε) pairs: (45, NUMBER), (year, TIME_PERIOD), (old, TIME_PERIOD), (africanamerican, Persons), (woman, Persons), (hormonereplacementtherapy, Therapeutics) emphasizing specific domains in terms of children of sixteen main branches of MeSH.(45, NUMBER), (year, TIME_PERIOD), and (old, TIME_PERIOD) remained unchanged and (african, ANTHROPOLOGY) and (american, GEOGRAPHY) are replaced with (africanamerican, Persons). (woman, FACTOTUM) is replaced with (woman, Persons). (hormone, ANATOMY), (replacement, FACTOTUM), and (therapy, MEDICINE) are replaced with (hormonereplacementtherapy, Therapeutics). Thus, four domains, NUMBER, TIME_PERIOD, Persons, and Therapeutics, are generated from an original text “45 year old african american woman with hormone replacement therapy”.

WordNet describes a word “45” as “a cardinal number” and WordNet Domains defines it as a domain NUMBER. However, we notice that “45” is not just “a cardinal number” or NUMBER but “the age of person” because it is used with other words in context; we can estimate a specific meaning for “45” from words “year old african american woman” including domains: TIME_PERIOD and Persons. DF algorithm estimates the meaning of certain domains by involving human’s intension. For example, if NUMBER is followed by two domains:

TIME_PERIOD and Persons, we can state that NUMBER is “the age of person” in context. We formallydefine a way to refine domains in the final stage of DF algorithm.

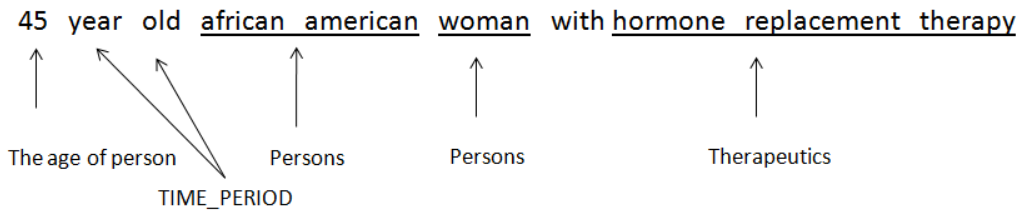

Figure 4.4 Word-domain pairs generated by DF algorithm

Figure 4.4 shows an example of (w, ε_f) pairs: (45, The age of person), (year, TIME_PERIOD), (old, TIME_PERIOD), (africanamerican, Persons), (woman, Persons), (hormone replacementtherapy, Therapeutics)generated by using DF algorithm. The age of person is from refined domains, TIME_PERIOD is from general domains, and Persons and Therapeutics are from specific domains. (w, ε_f) pairs will be used for a statistical topic model in Section 4.4.

We define Uas a refined domain set which consists of two subsets:$U_p = \{\varepsilon_1, \varepsilon_2, \dots, \varepsilon_p\}andU_u = \{\varepsilon_1, \varepsilon_2, \dots, \varepsilon_u\}$, where $U_u$is anuser-defined domain set created by a domain user manually and $U_p$isa pre-defined domain set from existing domains. An elementin $U_u$is substituted for one or more elements in $U_p$when it meets the rules defined by the domain user. A function: $I_P: X \rightarrow \{0,1\}$, where I_P indicates whether pre-defined domains are in a window or not, is defined as:

$$I_P(\varepsilon) = \begin{cases} 1 \text{ if } \forall\varepsilon\{\varepsilon \in U_p \rightarrow \varepsilon \in X_{sub}\}, \\ 0 \text{ otherwise} \end{cases} \quad (4.3)$$

, where $X_{sub}$is a subset of X. Algorithm 4.2 summarizes DF algorithm.

Algorithm 4.2: Domain Fusion

Input: $\mathcal{W}, \mathcal{L}, \mathcal{G}, \mathcal{S}$, U_p, U_u, N

Output: $S_F = \{(w_1^f, \varepsilon_1^f), (w_2^f, \varepsilon_2^f), \ldots, (w_n^f, \varepsilon_m^f)\}$

1. $\mathcal{G}_{sub} = \emptyset$, $\mathcal{S}_{sub} = \emptyset$, and N = 0
2. foreach w, w $\in \mathcal{w}$ do
3. $\mathcal{G}_{sub} \leftarrow (w, \varepsilon_g)$
4. if ε_s level $\leq \mathcal{L}$
5. $\mathcal{S}_{sub} \leftarrow (w, \varepsilon_s)$
6. end
7. end
8. N = N + 1
9. T = $\emptyset$, $\mathcal{F}_{sub} \leftarrow \mathcal{G}_{sub} \cup \mathcal{S}_{sub}$ with a priority of ε_s
10. $\varepsilon \in X$, $(w, \varepsilon) \in \mathcal{F}_{sub}$
11. while $I_p(\varepsilon)$ do
12. foreach X do
13. $Y \leftarrow X, y \in Y$
14. if $I_p(y)$
15. $T \leftarrow \varepsilon_u$
16. else
17. $T \leftarrow \varepsilon_p$
18. end
19. end
20. X = T
21. end
22. $S_F \leftarrow X$
23. repeat 1 to 22 until $|\mathcal{W}| >$ num
24. returnS_F

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

4.4 Domain specific LDA model

In this section, we explain how (w, ε) pairs can be applied to LDAwith word meanings.

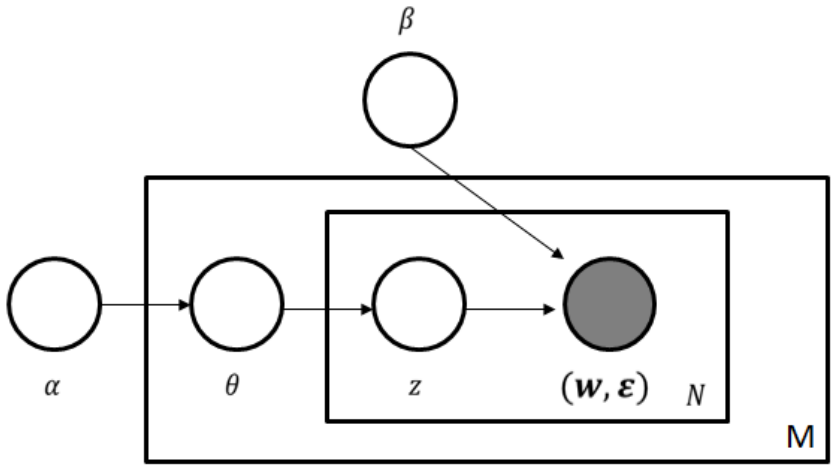

Figure 4.5DS-LDA representation

Figure 4.5 shows our graphical representation for DS-LDA. Each circle node indicates a random variable and the node shaded indicates (w, ε) pairs which are the only observed variables. Each plate represents replicates. Our model representation follows LDA model but the node shaded in the original LDA is substitute with (w, ε) pairs. The definition of terms follows:

- M: Number of documents
- N: Number of (w, ε) pairs in a document
- α: A corpus level parameter of the Dirichlet prior on the per-document topic distribution
- β: A corpus level topic$(z) \times (w, \varepsilon)$ pair matrix
- θ: A document level topic distribution; k-dimensional Dirichlet random variable
- z: A word level topic variables; k-dimensional multinomial random vector
- (w, ε): (word, domain) pairs

DS-LDA follows a generative process that considers hidden variables or hidden parameters to explain observed groups. Traditionally, in probabilistictopic models such as PLSA

and LDA, documents are represented by mixture of topics and a word w is followed by a topicz. Thus, we can find $p(z|(w,\varepsilon))$ learning $p(z)$and $p((w,\varepsilon)|z)$: $p(z|(w,\varepsilon)) \propto p(z)p((w,\varepsilon)|z)$.

Given $\alpha$and β, the joint distribution of θ, z, and (w,ε) pairs follows:

$$p(\theta, z, (w,\varepsilon)|\alpha,\beta) = p(\theta|\alpha)\prod_{n=1}^{N} p(z_n|\theta)p((w,\varepsilon)_n|z_n,\beta) \quad (4.4)$$

DS-LDA computes the posterior distribution $p(\theta, z|(w,\varepsilon),\alpha,\beta)$ to find the hidden variables $\theta$and z. However, the distribution is also intractable like the original LDA because of the coupling between θ and β when computing $p((w,\varepsilon)|\alpha,\beta)$.

$$p(\theta, z|(w,\varepsilon),\alpha,\beta) = \frac{p(\theta, z, (w,\varepsilon)|\alpha,\beta)}{p((w,\varepsilon)|\alpha,\beta)} = \frac{p(\theta|\alpha)\prod_{i=1}^{N} p(z_n|\theta)p((w,\varepsilon)_n|z_n,\beta)}{\int p(\theta|\alpha)(\prod_{i=1}^{N}\sum_{z_n} p(z_n|\theta)p((w,\varepsilon)_n|z_n,\beta))d\theta} \quad (4.5)$$

Thus, we perform approximate inference in DS-LDA model using the collapsed Gibbs sampling method. Gibbs sampling constructs a Markov chain computing the conditional distribution,$p(z_i|z_{-i},(w,\varepsilon))$, where z_{-i} represents the topic assignments for all (w,ε) pairs except $(w,\varepsilon)_i$. The conditional distribution is given by:

$$p(z_i = j|z_{-i},(w,\varepsilon)) \propto \frac{n_{-i,j}^{((w,\varepsilon)_i)}+\beta}{n_{-i,j}^{(\cdot)}+W\beta} \times \frac{n_{-i,j}^{(d_i)}+\alpha}{n_{-i,\cdot}^{(d_i)}+K\alpha} \quad (4.6)$$

, where $n_{-i,j}^{(d_i)}$is the number of (w,ε) assigned to topic *j* in document $d_i$excluding $(w,\varepsilon)_i$. $n_{-i,\cdot}^{(d_i)}$is the total number of (w,ε) in document d_i excluding $(w,\varepsilon)_i$. $n_{-i,j}^{((w,\varepsilon)_i)}$is the number of (w,ε) assigned to topic *j* excluding $(w,\varepsilon)_i$. $n_{-i,j}^{(\cdot)}$is the total number of (w,ε) assigned to topic *j* excluding$(w,\varepsilon)_i$. Thus, the first fraction represents the probability of $(w,\varepsilon)_i$ with a topic *j*and the second fraction represents the probability of a topic *j*in a document d_i.

4.5 Summary

In this chapter, we proposed a domain specific topic model that combines domain concepts with LDA. Two domain specific algorithms are introduced to generate domain concepts from document collections. These domain concepts are combined with LDA identifying the meaning of words.

5 MEDICAL DOCUMENT RETRIEVAL AND CLASSIFICATION WITH DOCUMENT SPECIFIC TOPIC MODEL

In this chapter, we propose new medical document retrieval and classification methods with our domain specific topic model. A query expansion method based on the domain specific topic model is proposed for the medical document retrieval. In addition, we describe how domains can be applied to state of the art classification models.

5.1 Background and problems

Recent advances in web and information technologies have resulted in dramatic increase in medical documents. Many approaches to handle these documents have been proposed to either complement existing techniques or make a technological breakthrough [54-56]. In the area of information retrieval, the recent technical issues aremainly dedicated to the usage of domain knowledge, such as genes, proteins and diseases [57, 58].

Using domains that covers a particular field of knowledge in information retrieval can be beneficial in concept representation of specific topics[59, 60]. However, information retrieval techniques that use only one concept may be limited by a narrow range of domain knowledge. For example, medical documents related to health disparities may contain a wide range of topics such as particular race or ethnics, relevant universities and regions, but a specific domain alone may not be useful to cover all of the topics because of its specialized characteristics in medical

documents.Moreover, the meaning of words can affect understanding of medical documents[61, 62]. Many approaches to identify the meaning of words have been mainly presented by using definitions in a dictionary [63-66] or by applying a statistical model [21, 23, 24, 36]. However, these approaches based on generalized terminologies of a dictionary are often inappropriate for medical documents due to the fact that the medical documents include specialized terminologies which may not be covered by traditional dictionaries. Therefore, they still have challenges with regards to the problems that involve understanding the word meanings in medical documents.

The meaning of words in context has been identified by determining word sensesdescribed in a dictionary, but they usually exist in a glossary form which is not be suitable for the use of real applications. To avoid this limitation in terms of word senses, some researchersproposed word sense identification methods that extract domain terminology from the word senses [9, 10]. The basic idea of the researches is that glosses are determined by its context, mapping theminto certain domains. These approaches have something in common with a concept of ontology. Ontology is a specification of a conceptualization that provides a formal frame representing a specific knowledge with a domain. We will adopt this ontology concept in the domain knowledge so that domains are conceptualized on multi-levels.

Domains extracted by word senses can be applicable to the fields of both information retrieval and text mining. As we already described in Chapter 2, varioustheoretical models such as Vector space model [3], Latent Semantic Indexing [4], Probabilistic Latent Semantic Indexing [5], and Latent Dirichlet Allocation [6]have been suggested to enhance the performance of information retrieval on many applications. However, their works on the applications have been presented by only dealing with pure text without any consideration of the meaning of words. This is because that they have primarily focused on creating new models to enhance retrieval

performance [67, 68]. In this chapter, we describe how domains can be applied to a query expansion, an application of information retrieval,using a topic model. We also show how domains can be used in the area of text mining, 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. Variousmodels have been proposed to increasethe effectiveness of the models [69-72]. However, these models mainly focus on finding optimum patterns of pure text with the limitation in terms of that the models often ignore the meaning of words. Some researchers have introduced various text mining models related to word senses, but their works are mainly focused on disambiguating word senses using their algorithms [73, 74]. In this chapter,we explain how domains can be applied to these models identifyingthe meaning of words with domains and showingthe effectiveness of the use of domains.

5.2 Our solution to the problems

In order to solve the problems described in Chapter 5.1, we propose a medical document retrieval method using a domain specific topic model. In addition, we show how domains can be applied to medical document classification models. Algorithm 4.1 and Algorithm 4.2 described in Chapter 4 are used for identifying specific domains in medical documents, and then three domains (WordNet Domains, MeSH, and Health Disparity) are applied to the proposed solutions.WordNet Domains is used to extract general domains that provide broad domain concepts, while specific domains: MeSH and Health Disparity, provide particular domain concepts that cover special domain knowledge. The overall solutions are described in [138].The main contributions of this chapter are as follows:

- Our approach takes word meanings into account when discovering domain knowledge from medical documents. Word senses in context are determined by the

proposed algorithms mappingthem to domains, and domains are extracted from the medical documents.

- Domain specific topic model is applied to a medical document retrieval method, which not only narrows domain concepts from different domainsbut also avoidsan unknown domain problem. Domainsareusedfor medical document classification, whichincreases the accuracy of classification models by identifying domains in a series of words.

5.3 Domain information

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

5.3.1 WordNet Domains

WordNet presented by G. Miller et al. [75, 76]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.As we described in Chapter 4, a set of synonyms called Synsetis a basic unit of WordNet andeach Synset can include 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 [77]. 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 byL.Bentivogli et al. [52], 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 [53] 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.

5.3.2 Medical Subject Headings

Medical Subject Headings (MeSH) is a controlled vocabulary thesaurus developed by National Library of Medicine (NLM) [78]. MeSH provides a hierarchical structure that covers several domains such as medicine, nursing and health care systems, consisting of headings in the twelve-level hierarchy. Thus, it has been mainly used for indexing biomedical articles or searching medical documents as well as retrieving meaningful text from documents [14, 15]. In 2014, it contains 27,149 descriptors and 218,000 entry terms indicating appropriate headings.

We use 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”.

5.3.3 Health Disparity Domains

Health Disparity (HD) refers to differences between groups of people with different races, ethnics and socioeconomics [79]. 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), 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.

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.

5.4 Experiments

In order to determine the effectiveness of the proposed model in medical documents, we conducttwo experiments based on Query Expansion (QE) and Text Classification (TC).

5.4.1 Query expansion

Query Expansion (QE) is a representative technique of information retrieval, which generates alternative queries on either lexical or semantic levels [13]. A variety of QE models have been proposed to enhance the effectiveness of information retrieval [17, 18, 80, 81], and it still has a great attention of many information retrieval communities today.

We describe how our proposed model is applied to QE. The proposed method has two advantages. First of all, we do not use sense definitions when expanding queries because they may cause a duplicated word problem when expanding queries. Instead, we use domains that contain refined concepts avoiding the redundant word problem. Second, hypernyms and synonyms are refined by topic words generated by domains.Note that the use of both hypernyms and synonyms without constraints such as levels, numbers, and ranges can degrade the performance of QE. Our method consists of four steps:

Step 1. Identify domains in document collections

Step 2. Generate topics from the Step 1

Step 3. Expand queries based on domains

Step 4. Remove domainswhich are not relevant to topics

First of all, we find domains in a set of documents. Because the purpose of our experiment is to verify the effectiveness of the proposed model, we initially identify the domains from words in the documents. As we described in previous chapter, Algorithm 4.1 and Algorithm 4.2 are performed based on both domain relevance and domain fusion.

Second, we generate topics from documents. A topic in a given documents can be represented by a set of words that shares same topics. These words can beusedas expanded queries for QE. Based on this concept, we will expand queries in the next step. To generate topics from the documents we use the proposed DS-LDA described in Chapter 4. Based on the

conditional distribution given by (4.1), we generate topic words from the result of the equation. The topic words generated by (4.1) will be used to remove unrelated words from the expanded query in the fourth step.

Third, we use domain 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 $\mathrm{p}(w|\mathrm{Q})$, where $w$is a word and Q is a query. Thanks to the theoretical foundation of information retrieval, we are able to estimate $\mathrm{p}(w|\mathrm{Q})$ in document aspect using$\mathrm{p_d}(w|\mathrm{Q}) = \sum_{\mathrm{D}\in\mathrm{C}} \mathrm{p}(w|\mathrm{D})\mathrm{p}(\mathrm{D}|\mathrm{Q})$, where D is a document and C is a set of documents. We define $\mathrm{p}(w|\mathrm{Q})$ for topic aspect:

$$\mathrm{p_t}(w|\mathrm{Q}) = \sum_{\mathrm{T}\in\mathrm{S}} \mathrm{p}(w|\mathrm{T})\mathrm{p}(\mathrm{T}|\mathrm{Q}) \qquad (5.1)$$

, where S is topics and T is topic words in S. By Bayes rule, $\mathrm{p(T|Q)} = \frac{\mathrm{p(Q|T)p(T)}}{\mathrm{p(Q)}} \propto \mathrm{p(Q|T)p(T)}$. We estimate $\mathrm{p_t}(w|\mathrm{Q})$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 conductedon OHSUMED[3] 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 chapter, 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 cosine similarity is computed by:

$$\mathrm{sim}(q,d) = \frac{q \cdot d}{|q||d|} = \frac{\sum_{k=1}^{n} q_{w_k} \times d_{w_k}}{\sqrt{\sum_{k=1}^{n} q_{w_k}^2} \times \sqrt{\sum_{k=1}^{n} d_{w_k}^2}} \quad (5.2)$$

, 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.

[3] http://trec.nist.gov/data/t9_filtering.html

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) [26]: 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) [18]: Queries are expanded by using lexical-semantic relations. Hyponyms are added to the original query from synonyms.
- Random Indexing (RI) [19]: 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 [26], 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 [82]. 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:

$$\mathrm{DCG_d} = \mathrm{count_1} + \sum_{i=2}^{d} \frac{\mathrm{count_i}}{\log_2 i} \tag{5.3}$$

$$\mathrm{nDCG_d} = \frac{\mathrm{DCG_d}}{\mathrm{IDCG_d}} \tag{5.4}$$

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

Figure5.1 shows the experimental results for $\mathrm{DCG_n}$. X-axis denotes accumulated DGG and y-axis denotes the retrieved document numbers. The result shows that DSS-LDA outperforms other methods from $\mathrm{DCG_{10}}$ to $\mathrm{DCG_{50}}$. In particular, the increase rate of $\mathrm{DCG_n}$ in DSS-LDA is larger than other methods and this explains the search performance of DSS-LDA is better than others.

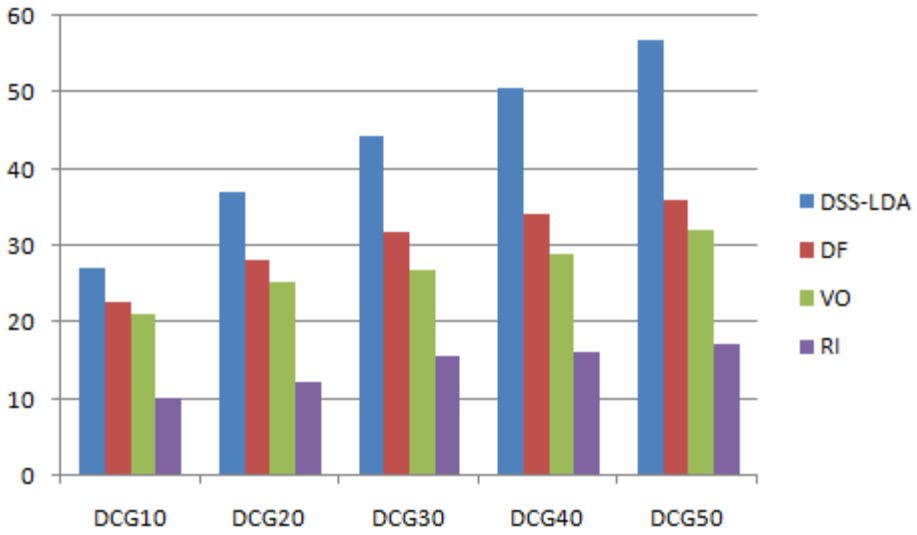

Figure 5.1DCG comparison of four models

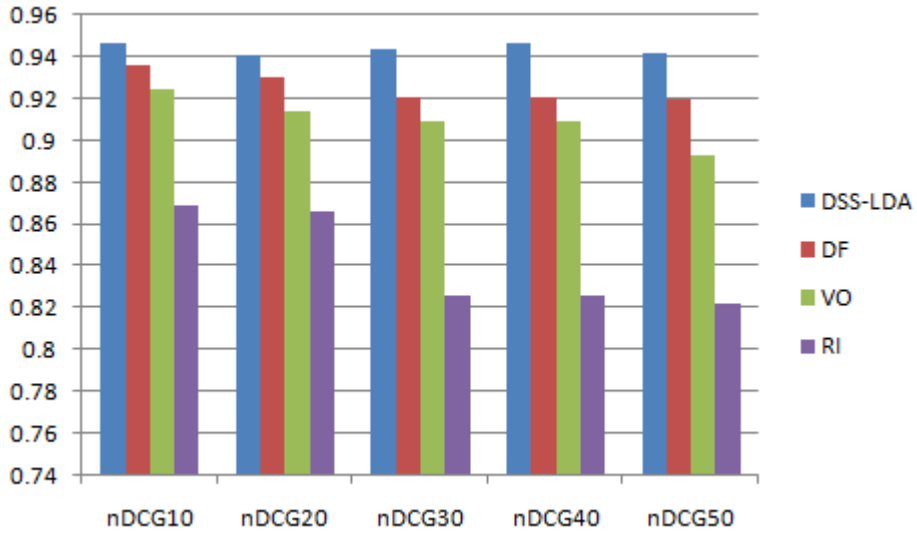

Figure 5.2 nDCG comparison of four models

Figure5.2 shows the experiment result for $nDCG_n$. X-axis denotes accumulated nDCG and y-axis denotes nDCG value ranges from 0 to 1, meaning that nDCG is a perfect value when it is 1. The overall results show that DSS-LDA is very good in the all $nDCG_n$ performance. In particular, DSS-LDA also has good results in $nDCG_n$ where n 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.

5.4.2 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: J48, NBTree, NaïveBayes and LibSVM, are used for evaluating the effectiveness of domains uses. J48 is a Java implementation of C4.5, a decision tree algorithm [83] and NaïveBayes is a well-known supervised learning algorithm that applies Bayes' theorem

[84]. NBTree is a hybrid version of a decision tree and naïve Bayes that generates a decision tree at the leaves [85] and LibSVM is an open source tool supporting Support Vector Classification (SVC) [86]. We use WEKA [87], 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.

Table 5.1 Accuracy for 6 groups of documents

Classifier	Domain	NCI	NEI	NHLBI	NHGRI	NIAID	NIMH
J48	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

Table 5.1 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.

Figure 5.3 F score comparison of 6 sets of documents

Figure 5.3 shows the experimental results for F-score. 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 algorithms: J48 and NaiveBayes produce the opposite results compared with J48 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 5.2 illustrates the performance comparison between classifiers with or without domains. According to Accuracy of the four classifiers, J48 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.

Table 5.2Accuracy for NIMHD

Domain	J48	NBTree	NaiveBayes	SVM
With	**0.935**	0.775	**0.895**	**0.9**
Without	0.8	**0.81**	0.665	0.5

Figure 5.4 shows Precision, Recall, and F1 scores for NIMHD. The best F1 score is J48 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.

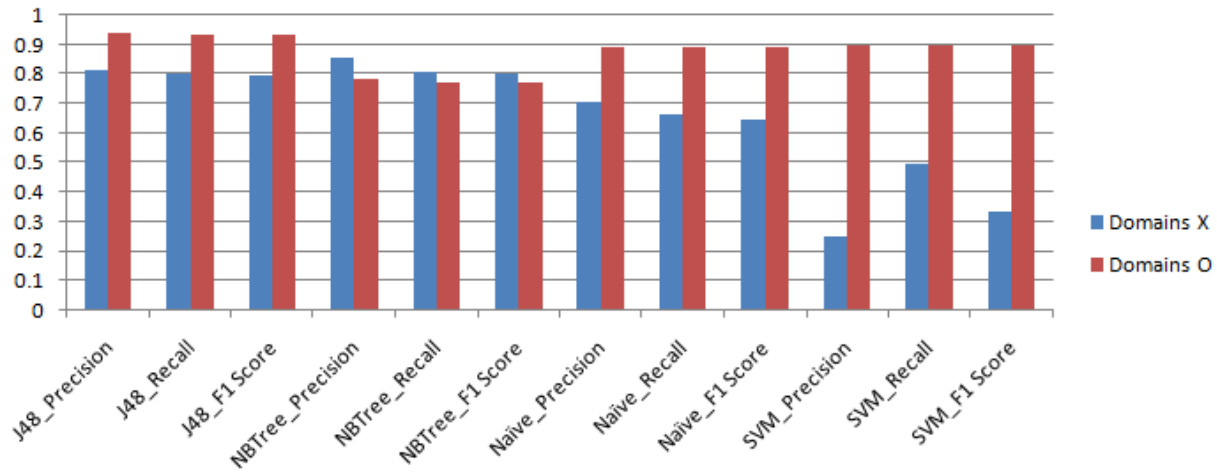

Figure 5.4Experimental results for Precision, Recall, F-score

5.5 Summary

We described how domains can be applied to two research areas: information retrieval and text classification. Experiments were conducted on a query expansion technique using the proposed model as well as text classification models using domains. The experimental results showed that the proposed model outperforms others in information retrieval, and domains can be very useful for text classification.

6 DOCUMENT SUMMARIZATION METHOD WITH DOMAIN SPECIFIC TOPIC MODEL

In this chapter, we propose a novel document summarization method that uses domain specific topic model. We applied the domain specific topic model to a document summarization to increase the effectiveness of the summarization method. The proposed method is compared with traditional summarization methods.

6.1 Background and problems

Dramatic changes in recent technologies have engaged people's attention to massive information resources that requires new strategies for summarization. Automatic summarization is a well-established research area that summaries the large volume of information automatically in a smaller one that retains essential information providing new observations. It has received much attention recently because of its ability to produce a condensed version of social media contents.

Various studies have beenconducted to improve the quality of summary in the social media. Sharifi et al. [88] proposed a phrase based algorithm that uses trending phrases for summarizing micro-blogs and Inouye [89] presented a multi-post summarization method that consists of two algorithms: a clustering algorithm and a threshold algorithm, to increase the effectiveness of the summarization. These methods have been compared with traditional summarization methods: MEAD [90], LexRank [91], and TextRank [92], and Inouye [93] also compared them with SumBasic [94]showing that SumBasic produces the best F-measures on Twitter. Zhang et al.[95]proposed a speech act-guided summarization method that focuses on speech acts of tweets.

However, these studies focused on pure text only might be vulnerable to various aspects such as a posting time, meaning of words and unique factors related to characteristics of social media. Also, tweets have some unique characteristics: short length of text messages, hashtags, and followers. However, the unique characteristics of tweets are neither fully considered nor integrated in the previous studies.

6.2 Our solution to the problems

In order to solve these problems, our method takes three aspects into consideration. We propose a tweet scoring method considering four different unique factors: tf-idf, tweets length, hashtag relatedness, and delivery weights. We will describe the details in the following chapter.

The rest of the chapter is organized as follows. In Section6.3,we describe our strategies for summarization. Two experiments are performed to determine the strength of the proposed method in Section6.4. Finally, conclusions will be given in Section6.5.

6.3 Text summarization with multi-aspects

Building summaries on tweets is to arrange tweets in order of importance. We propose a method of scoring importance weights computed by combinations of four different factors: term frequency - inverses document frequency (*tf-idf*) with DSTM, tweet length (*tl*), hashtag relatedness (*hr*) and delivery weights (*dw*). We define total scores as below:

$$S = S_{tf-idf} \times S_{tl} \times S_{hr} \times S_{dw} \tag{6.2}$$

tf-idf is a well-known method that has been used to measure the importance degree of a sentence or a document. Our DSTM is used to generate pairs of word and domain from original

documents, and then used for the method. *tf* is used to measure the word frequency in a tweet and *idf* is used to measure the tweet frequency. We define *tf-idf* score as below:

$$TF - IDF = TF(f, e) \times IDF(f)$$
$$S_{tf-idf} = \frac{\sum_{i=1}^{n} TF - IDF_i}{(\sum_{i=1}^{n} TF - IDF_i)_{\max}} \quad \text{for } e \in E,\ f \in F_{fusion} \qquad (6.3)$$

, where *TF(f, e)* is a word frequency in a tweet *e* and *IDF(f)* is an inverse tweet frequency in a set of tweets *E*. *n* is a total number of words.

A length of tweets may affect the importance of tweets. Traditional approaches for document summarization assume that the shorter the document length, the better the document importance is. *tl* is taken into account because users in Twitter tend to oversimplify tweets in which messages are short and clear. We compute the *tl* score by normalizing it given total number of words in a tweet. The score follows:

$$S_{tl} = \frac{L_e}{(L_e)_{\max}} \qquad (6.4)$$

, where L_e is the length of words in a tweet *e* and *max* indicates the maximum.

Hashtags used by adding # to a tag have been an effective way of organizing topic information on Twitter. With the help of the hashtags, people are able to post a tweet indicating certain topics or issues more easily. We consider the hashtags as an important factor for our summarization method. The *hr* is computed by the below:

$$S_{hr} = \begin{cases} 0 & \text{for } |h|=0 \\ 1 & \text{for } |h|=1 \end{cases}$$

$$\frac{\sum_{i=1}^{n} sim_i}{|h|} \quad \text{for } |h|>1 \tag{6.5}$$

, where $|h|$ is the number of hashtags in a tweet and sim_i is a degree of similarity between hashtags.

The number of followers or fans on Twitter may affect the importance of tweets. For example, Figure 6.2 shows that the number of tweets is proportional to the number receivers. We impose weights on tweets considering the aspect named *dw*. We define *dw* as the below:

$$S_{dw} = \frac{|f_u|}{sum_t(f)} \text{ for } u=1,2,\ldots,n \tag{6.6}$$

, where $|f_u|$ is the number of followers for a user *u* and $sum_t(f)$ is the sum of all followers in a time *t*.

6.4 Experiments

In this section, we present experimental results for the proposed summarization method. We introduce a real dataset collected to be used for the experiments from Twitter. Second, we explain about Recall Oriented Understudy for Gisting Evaluation (ROUGE) [101], a well-known evaluation metric in the field of automatic summarization. We show the experimental results comparing the proposed method with other traditional summarization methods.

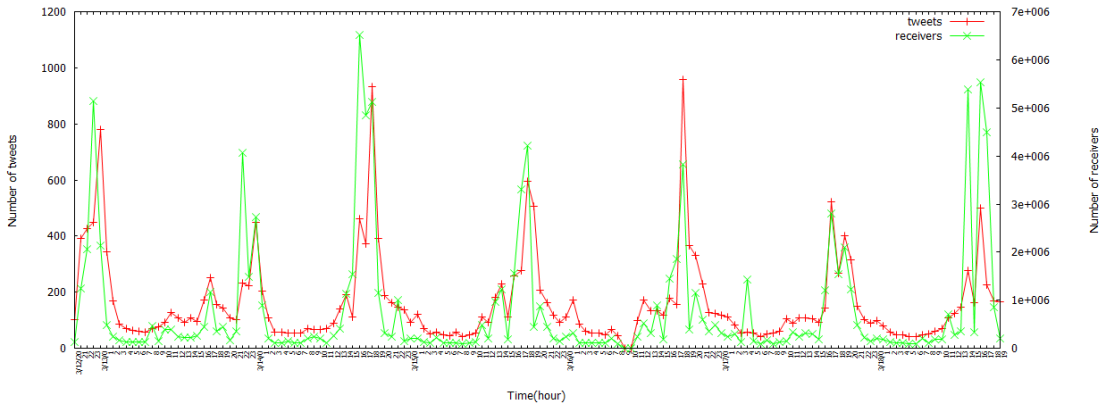

Figure 6.1 Tweets and receivers collected from Twitter

6.4.1 Dataset

Figure 6.1 shows the data variation aboutthe number of receivers in a period of time. Tweets from Twitter which is one of most popular social media with a hashtag, #Dodgers, are collected during a week, from 3/12/2015 8:00 PM to 3/18/2015 7:00 PM. Tweets indicate a short message written by a person and a receiver represents someone who follows the tweets. The follower counts of tweets can be considered as numbers of receivers. Both tweets and receivers in a period of time give us a new idea about how topics with hashtags catch people's attention and how many tweets are delivered to people affecting the topics.

We use a hashtag, #Dodgers, collected from Twitter public streaming data for 7 days from 3/12/2015 8:00 PM to 3/18/2015 7:00 PM. A week is enough time for our experiments to perform the summary evaluation since the streaming data has been entered in every millisecond. In particular, for #Dodgers, it is a suitable time range to receive Dodgers game information because it plays 7 times a week. Total 25,191 tweets are collected with 114,755,852 receivers expected during 144 hours. Four graduate students have participated in summarizing the tweets manually. We ask the volunteers to summarize all tweets so that model summaries consist of three different summaries. The model summaries are compared with system summaries

generated by five different summarization methods: *mats*, *fs* [96], *ots* [105], *swe*[106], *tf-idf*. The *fs* is a Twitter summarization mehtod that uses fuzzy-inference system [107] to extract important sentences from tweets in real-time. The *ots* is a freely availabe summarization tool used as a benchmark for many text summarization methods [108-110]. The *swe* is an automatic summarization method focused on language independent summarization that has been evaluated on large-scale dataset. The *tf-idf* is a well-known vector space model used as a baseline for summarization. Our summarization method consists of tweets selection, domain centered word sense identification and tweets scoring which is the combination of *tf-idf*, *tl*, *hr* and *dw*. We call this as Multi-Aspects Twitter Summarization (*mats*) throughout this chapter. Baseball-ont[4], a baseball ontology, is used for the #Dodgers domain information. We also used protégé[5], a well-known ontology editor, to build #Dodgers domain information following the structure of Baseball-ont. WordNet Domains[6] is used to extract other domain information.

6.4.2 Evaluation metric

To measure the effectiveness of summarization methods we adopt ROUGE metric commonly used in summarization evaluation. ROUGE metric has been widely used for summary evaluation. The metric enables comparing performance in different systems on the same set of documents, assuming that model summaries are available for those documents. We compare *mats* with other summarization methods based on ROUGE-N metrics which is an n-gram recall between system summaries and model summaries. The term n-gram denotes a sequence of n successive words and n stands for the length of n-gram. The ROUGE-N is computed by counting the number of overlapping words between system summaries and model summaries. ROUGE-N is defined as below:

[4] http://www.daml.org/2001/08/baseball
[5] http://protege.stanford.edu
[6] http://wndomains.fbk.edu

$$ROUGE - N = \frac{\sum_{S \in M} \sum_{gram_n \in S} N_{match}(gram_n)}{\sum_{S \in M} \sum_{gram_n \in S} N(gram_n)} \tag{6.7}$$

,where *M* is model summaries and *n* is the length of the n-gram. $N_{match}(gram_n)$ is the largest number of n-grams that co-occurs in a system summary and a set of model summaries.

In our experiment, we use ROUGE-1, ROUGE-2, ROUGE-3, ROUGE-4, ROUGE-L, ROUGE-W-1.2, ROUGE-S and ROUGE-SU metrics. ROUGE-L is the longest common subsequence (LCS) based statistics that finds the longest common subsequence of two sequences of items. Lin et al. [102] presented the LCS evaluation between a system and a set of model translation. ROUGE-W-1.2 is weighted longest common subsequence with the weight of 1.2. Since the basic LCS has a problem that LCS does not consider spatial relations within sequence, we also use the ROUGE-W- 1.2 metric. ROUGE-S is based on skip-bigram, a pair of words in sentence order and ROUGE-SU is added unigram-based co-occurrence statistics to the skip-bigram. These metrics generate the recall, precision, and F-measure scores. The scores are defined as below:

$$\begin{aligned} \mathrm{Re}\, call(s \mid m) &= \frac{|\, s \cap m \,|}{m} \\ \mathrm{Pr}\, ecision(m \mid s) &= \frac{|\, s \cap m \,|}{s} \\ F - measure &= 2 \times \frac{precision \times recall}{precision + recall} \end{aligned} \tag{6.8}$$

, where *s* indicates system *m* indicates model. The intersection of system summaries and model summaries is the number of words that the summaries shares.

6.4.3 Experiment results

In this section, we present our experimental results on the ROUGE-1. For the experiment, five summarization methods: Multi-Aspects Twitter Summarization (*mats*), Fuzzy Summarization (*fs*), Open Text Summarization (*ots*) and Term Frequency – Inverses Document Frequency respectively (*tf-idf*) are compared with each other.

Table 6.1 shows a week schedule for Dodgers and experimental time periods with λ: 100, 200 and 300. Los Angeles Time is converted into Eastern Daylight Time (EDT) because we have collected tweets based on EDT zone. For example, the time 3/14 16:05 on the schedule should be 3/14 13:05 but, for the convenience, we converted it into EDT. We can intuitively see that the selected time periods correspond to the real schedule of Dodgers game. We compare our proposed summarization method named mats with other methods using the tweets in the time periods.

Table 6.1 Schedule and selected time periods

λ_p, λ_n	100	200	300
3/12 16:05	3/12 20 - 3/13 1	3/12 21- 3/13 1	3/13 0 - 3/13 1
3/13 22:05	3/13 22 - 3/14 1	3/14 0 - 3/14 1	
3/14 16:05	3/14 16 - 3/14 19	3/14 16 - 3/14 19	3/14 16 - 3/14 19
3/15 16:05	3/15 16 - 3/15 20	3/15 18 - 3/15 20	3/15 18 - 3/15 20
3/16 16:05	3/16 11 - 3/16 19	3/16 18 - 3/16 19	3/16 18 - 3/16 19
3/17 16:05	3/17 17 - 3/17 21	3/17 17 - 3/17 18	
3/18 16:05	3/18 14 - 3/18 17	3/18 16 - 3/18 17	3/18 16 - 3/18 18

Figure 6.2 shows ROUGE-1 comparison of summarization methods where λ is 100. The x-axis illustrates the time periods and the y-axis illustrates F-measures of five different summarization methods scored at ROUGE-1 level. The result shows that *mats* outperforms other methods significantly on ROUGE-1. Figure 6.3 and Figure 6.4 show ROUGE-1 comparison with

different λ. In Figure 6.3, *mats* still outperform other methods and *swe* shows the best F-measure on 3/13. However, *swe* presents highly irregular results on many days. For example, F-measure is very high on 3/13 but it is very low on 3/12. Furthermore, the average of F-measure on ROUGE-1 is lower than *mats*. We will show this on Figure 6.5. There are no big differences in Figure 6.4, excepting on 3/15 and 3/16. We can see that the time periods on 3/14 and 3/15 are same when λ is 200 and 300 on Table 6.1. This means that tweets collected on the days are same with tweets used in Figure 6.3, indicating that two λ values are in the same time periods. There are no F-measures on 3/13 and 3/17 because tweets are not found when λ is 300. Figure 6.5 shows ROUGE-All averages of summarization methods with 95 % confidence interval. The x-axis illustrates all ROUGE metrics and the y-axis illustrates the averages of F-measures for ROUGE metrics. We notice that the averages of ROUGE-1 F-measures are higher than F-measures' averages of other metrics. This is because that people tweeting messages on Twitter write short messages of 140 characters so that a single word is more effective than multi words. We further notice that the averages of F-measures of all ROUGE metrics on the mats outperform others significantly. This illustrates that the averages of all ROUGE metrics of Precision of Recall on the mats are higher than others.

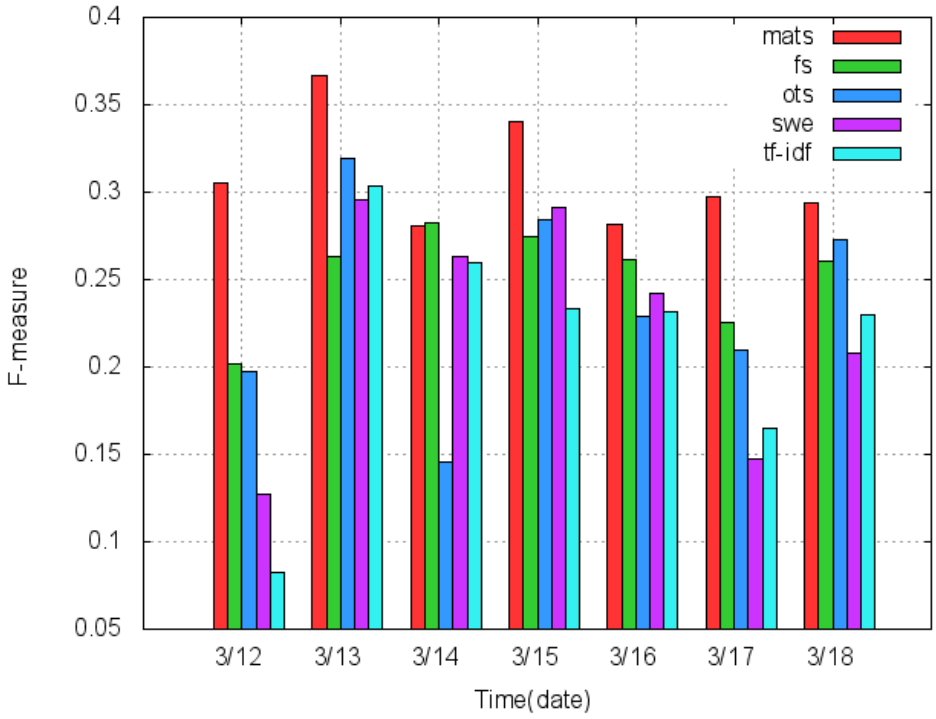

Figure 6.2ROUGE-1 comparison of summarization methods (lambda =100)

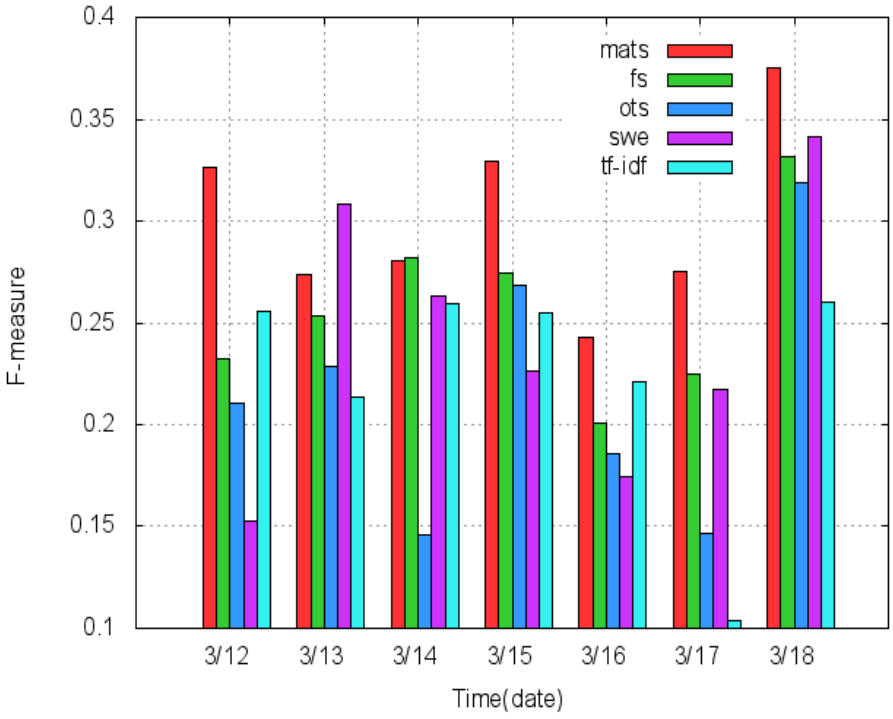

Figure 6.3ROUGE-1 comparison of summarization methods (lambda =200)

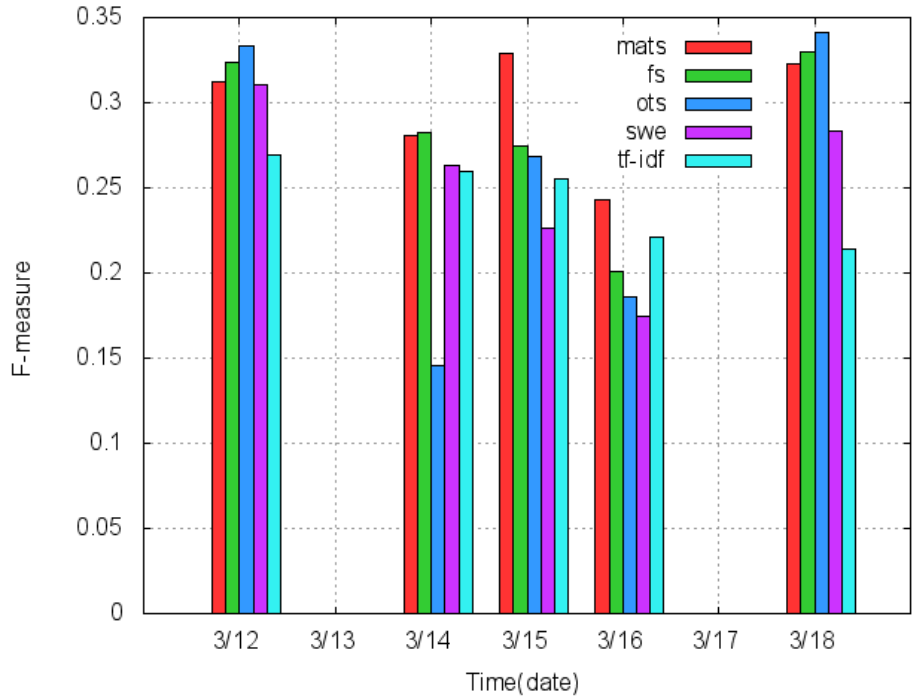

Figure 6.4ROUGE-1 comparison of summarization methods (lambda=300)

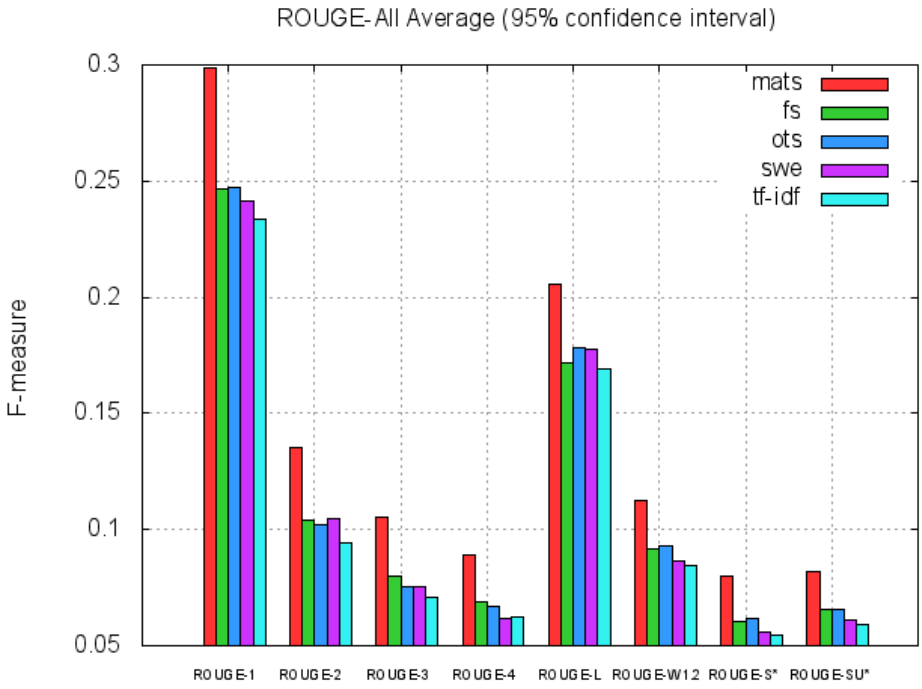

Figure 6.5ROUGE-All averages of summarization methods

6.5 Summary

We proposed a novel document summarization method using a domain specific topic model. A domain specific topic model is used to select important tweets from the period of time. Moreover, a tweet scoring method is presented to consider four different unique factors.

The experimental results showed that the proposed method significantly improves summarization performance when all aspects are applied to the summarization task. As a result, our summarization method outperforms traditional summarization methods on all aspects.

7 TAG BASED IMAGE RETRIVEAL METHOD WITH DOMAIN SPECIFIC TOPIC MODEL

In this chapter, we propose a tag-based image retrieval method that uses a domain specific topic model. A domain specific process on the proposed method is presented by combining a domain specific topic model for tags with low level features for visual contents.

7.1 Background and problems

According to Yahoo over 800 billion photos were taken by the users on the web in 2014 and will be grown exponentially every year. One of the reasons for this gigantic explosion of images is the popularity of smart-phones and digital cameras sharing photos on public web sites. Withthis new era of photography, tagging on social images has been also flourished and became a routine activity of the users [127-130].

Figure 7.1 Results for a query “airport” on Flickr

In particular, tags used to describe images have played an important role in enabling the users to search relevant images directly in social media. However, tags are often arbitrary words that are user dependant increasing a gap between a provider and a searcher. Generally, providers assign tags when uploading images on social media, while searchers predict tags when retrieving the images.The prediction of tags is more difficult for searchers who lack any prior knowledge of the images. This increases the gap between the intended meaning of providers and the searchers, decreasing an accuracy of image retrieval. An example of this problem is shown in Figure 7.1. A query “airport” can be results in many different images. The driving point for our motivation is to overcome this limitation in social media, increasing the quality of social image retrieval.

Many studies have been done enhancing the social image retrieval performance [103, 111-113, 131-134]. M. Wang et al. [114] and Y. Gao et al. [115] showed that a relevance degree between visual contents and textural information can be the best way of increasing image retrieval performance. X. Li et al. [117] maintained that users’ activities can affect on the social image retrieval performance. G. Zhu et al. [118] presented a new framework for a tag refinement

in a large volume of social images. They used both visual contents and textural information to increase the tag refinement and showed that their method is more effective than other contemporary methods. J. Sang et al. [116] proposed a tag refinement method that uses ternary relations on large scale images. They described a ranking based technique with user tagging behaviors and showed the effectiveness of the method using on [119]'s evaluation framework. However, their method does not consider the meaning of tags and is mainly focused on refinement of tags. J. Tang et al. [120] presented an image retrieval framework to reduce a semantic gap between low level features and high level concepts. They constructed a concept space that infers semantic concepts from community-contributed media including both images and tags, and then applied a graph-based learning method into the concept space. Their experiments were conducted on a light version of NUS-WIDE database [121] and showed its effectiveness, but their dataset size was relatively small. Y. Gao et al. [122] proposed a social image search method that uses both visual contents and tags on images. A hyper-graph was constructed by combining visual contents with tags, and then a relevance learning method was conducted on the hyper-graph structure. They showed that their method outperforms other approaches such as semi-supervised learning and tag ranking.

All of previous studies have mainly focused on combining low level features with textural information on images. However, they didn't use domains which are suitable for identifying the meaning of words in tags. We extract the meanings of the textural information, and then combine them with low level features. This is fundamentally different from other methods that only focus on improving image retrieval performance through combination of low level features and just given textural information. The main advantage of the proposed method is that our model can be applied to any kind of retrieval method that uses textural information.

7.2 Our solution to the problems

In order to reduce semantic gaps between a provider and a searcher, we propose a semantic processing using domains through a sequence of steps in which the meanings of tags is identified. In this chapter, we define the semantic processing as Domain Specific Semantic Process (DSSP) on tag-based image retrieval.

DSSP uses two domain specific algorithms proposed in Chapter 4 to identify the meaning of tags for images based on both WordNet Domains [77] and English Wikipedia entities [123]. The algorithms find domains for each tag and Latent Dirichelet Allocation (LDA) [6], a topic model introduced in Chapter 4, will generate topic distributions of tags with domains. And, we compute probabilities of the tags given similar images by exploiting a content-based image retrieval technique called Bag of Visual Words (BoVW), which provides relations between tags and images. The overall framework is illustrated in Figure 7.2.

Figure 7.2of the tag-based image retrieval with the proposed model

The contributions of this chapterare as follows: 1. we propose a semantic process that enables searchers to retrieve images with unrefined tags by adding domains to each tag. 2. we use two domain specific algorithms to narrow semantic concepts from broad domains to specific domains. 3. The proposed process includes the combination of tag-domain sets and LDA. This way allows us to find relevant tag-domains sets providing topic distributions.

This chapter organized as follows. Section 7.3 describes our proposed method in details. In Section 7.4 we will show our experimental results. Finally, conclusions and future works will be given in Section 7.5.

7.3 Tag-based image retrieval with domain specific topic model

In this section, we describe a new domain specific semantic process for tag-based images retrieval. Two domain concepts, WordNet Domains and English Wikipedia Entities, are explained in Section 7.3.1. In Section 7.3.2, we describe a relevance measure for tags and visual contents used in the proposed method.

7.3.1 Domain concepts

To identify domain concepts in tags we use WordNet Domains and English Wikipedia Entities. As we described in Chapter 4.1, WordNet Domains is a well-known lexical resource that is annotated by using WordNet with semantic domain labels.It is structured on the basis of 200 domains generated in a hierarchical structure semi-automatically. Each sense of a word is labeled with one or more domains and a label FACTOTUM is assigned for a special case of domain that is unknown.

Our proposed system is mapping textual strings to canonical URLs of English Wikipedia Entities [123]. A particular string can have various resources such as Wikipedia titles and links within the contents of Wikipedia. Table 7.1 shows resource entries for matching a string Hank Williams". In order to find closely related Wikipedia URLs, the conditional probabilities of URLs given a string s: S(URL|s) is used. Matching a string for finding entries may be considered as a typical entity linking task. In fact, tags in social media are usually exposed to a variety of entities which makes the task of mapping them to a particular entity very difficult in wide entity distribution. In this chapter, we plan to combine English Wikipedia Entities with WordNet Domains to generate the meaning of tags rather than using only particular entity recognition method. This is a valid reason for using the English Wikipedia concepts. WordNet Domains with the English Wikipedia Entities is combined by Algorithm 4.1 and 4.2.

Table 7.1English Wikipedia example

S(URL|s)	Canonical(English) URL
0.990125	Hank_Williams
0.00661553	Your_Cheatin'_Heart
0.00162991	Hank_Williams,_Jr.
…	…
0.0000958773	Hank_Williams_(basketball)

7.3.2 Relevance between tags and visual contents

Some tags may not be closely related to visual contents of image when tags have additional information not related to visual contents. In order to compute the relevance degree between tags and visual contents, we use k Nearest Neighbors (k-NN) from Bag of Visual Words (BoVW) [115, 121]. BoVW is a popular local feature based technique for content-based image retrieval inspired by Bag of Words (BoW) model. Typically, an image can be translated into a set of visual words using key points collected from images to describe salient regions and local features clustered to generate visual vocabularies. For example, in SIFT [124] descriptors, key points are often used as feature vectors. However, there is a need to overcome a limitation of correlating visual words with tags because BoVW only relies on the discriminative power of visual vocabulary.

To compute the relevance between tags and visual contents, we find a probability of tags given k-NN from BoVW. The relevance degree is computed by:

$$R_t = P(t \mid N_t(I_k)) \qquad (7.1)$$

, where t is a refined tag and $N_{\mathrm{t}}(I_{\mathrm{k}})$ is k-NN with t. The higher probability leads the higher relevance between tags and visual contents.

7.4 Experiment

In this section, experimental results will be shown to demonstrate the performance of the proposed method on a large-scale image dataset.

7.4.1 Dataset

NUS-WIDE [121] is a well-known social image dataset that includes 269,648 images and tags associated with the images. The dataset is collected from Flickr with 5,018 unique tags. We use 49 concepts with 500 dimensional bag of visual words based on SIFT descriptor, which are provided by NUS-WIDE. Figure7.3and Figure 7.4 show distributions of ground-truth of the concepts. X-axis indicates concept names and y-axis indicates the number of relevant images for the concept names.

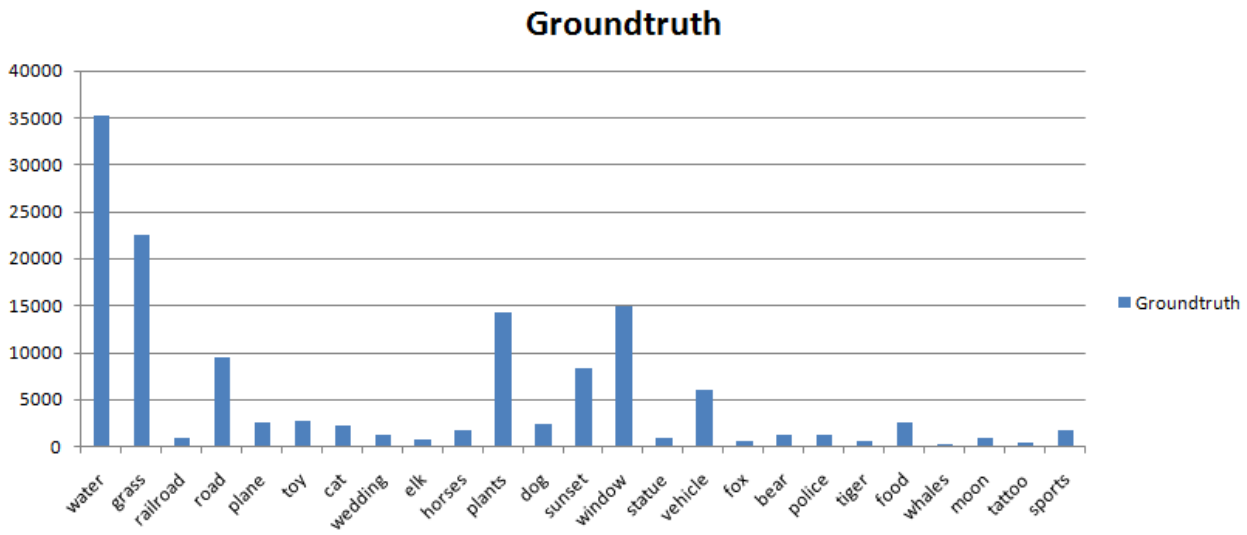

Figure 7.3Distribution of Ground-Truth of 25 concepts

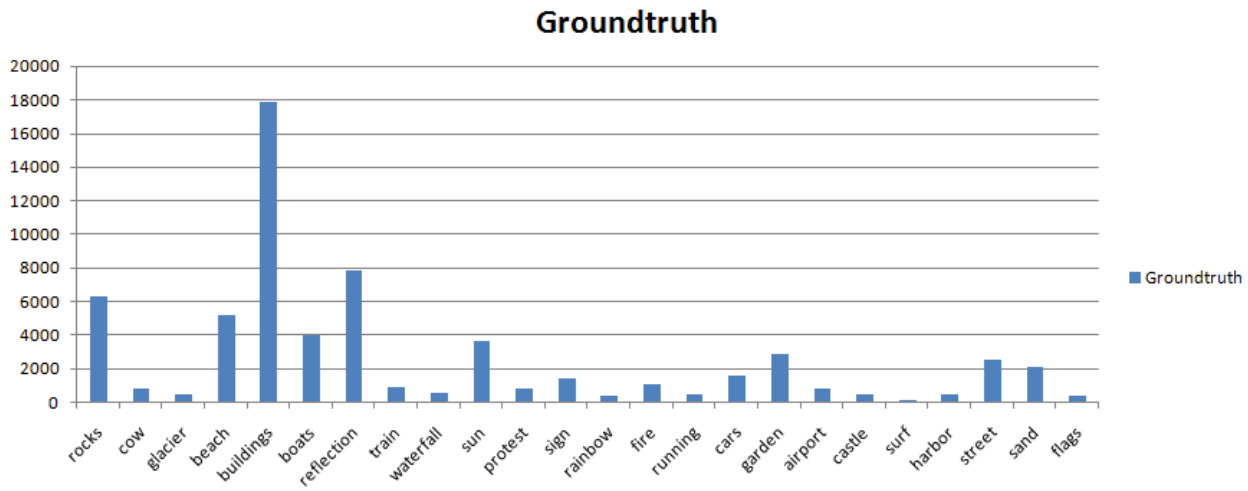

Figure 7.4Distribution of Ground-Truth of 25 concepts

7.4.2 Evaluation metric

Our experiment of the proposed method is conducted with two different aspects: one domain and multi-domains. First of all, we perform an experiment with one domain to demonstrate how the meaning of tags on social images affects Precision of keyword search results. Because the keyword search uses text only, we do not apply BoVW as well as any topic models to the experiment. This is because that the purpose of the first experiment is to verify the effectiveness of the use of domains. For example, when the user types a keyword "airport" as a query, the keyword search results will be a set of images that contain a tag "airport". On the other hand, when the user type a keyword "airport" as a query, our method will add a domain to the query so that "airport" will be "airport|TRANSPORT". Thus, the keyword search results will be a set of images that contain a tag-domain "airport|TRANSPORT". For the first experiment, we use 49 concepts extracted from social images providing general domain concepts of WordNet Domains. Note that we use only one domain for each concept because we assumed that 49 concepts are ideal and do not have multi-domains. Table 7.2 shows 49 concepts with the assigned domains. Next, we perform another experiment with multi-domain. Unlike the first

experiment that uses only one domain, the second experiment uses multi-domain. This is very important for us to do the second experiment because several concepts may have various meanings. For example, WordNet Domains defines "airport" as TRANSPORT. However, the meaning of the domain is too broad to identify every social image because there are many cases that the users tag "airport" on their social images which are not relevant to TRANSPORT. The "airport' can be related to other domains, such as UNDERCARRIAGE, AVIATION, and PERSON. Therefore, it is necessary for us to narrow the range of domain into more specific ones. In order to solve this problem, we use a refined domain set introduced in Chapter 4.

Table 7.249 concepts with domains

Tag	Domain	Tag	Domain	Tag	Domain	Tag	Domain
Airport	TRANSPORT	Food	FOOD	Rainbow	NATURE	Tattoo	ART
Beach	GEOLOGY	Fox	ANIMALS	Reflection	PHYSICS	Tiger	ANIMALS
Bear	ANIMALS	Garden	AGRICULTURE	Road	TRANSPORT	Toy	PLAY
Boats	NAUTICAL	Glacier	GEOLOGY	Rocks	GEOLOGY	Train	TRANSPORT
Buildings	BUILDINGS	Grass	PLANTS	Running	SPORT	Vehicle	TRANSPORT
Cars	TRANSPORT	Harbor	GEOGRAPHY	Sand	GEOLOGY	Water	CHEMISTRY
Castle	BUILDINGS	Horses	ANIMALS	Sign	TELECOMMUNICATION	Waterfall	GEOGRAPHY
Cat	ANIMALS	Moon	ASTRONOMY	Sports	SPORT	Wedding	RELIGION
Cow	ANIMALS	Plane	TRANSPORT	Statue	SCULPTURE	Whales	ANIMALS
Dog	ANIMALS	Plants	PLANTS	Street	GEOGRAPHY	Window	BUILDINGS
Elk	ANIMALS	Police	ADMINISTRATION	Sun	ASTRONOMY		
Fire	FLAME	Protest	SOCIOLOGY	Sunset	TIME_PEROID		
Flags	ART	Railroad	TRANSPORT	Surf	SURF		

Two experiments are conducted by preprocessing steps needed to refine tags in social images. Because the tags may contain unnecessary words such as 'a', 'the' and '-s', it is necessary to remove the tags in pre-processing steps. To do this, our experiments include both stop-word removing and stemming for all tags. For the second experiment, we follow DSSP presented in Figure 7.2. The experiment assumes that the meaning of images can be found on domain concepts. For example, if the users search an image that describes a man in airport, tag-domains will be "man|PERSON" and "airport|TRANSPORT". We use three domain specific concepts, UNDERCARRIAGE, AVIATION, and PERSON for "airport|TRANSPORT". The

proposed topic model generates k-specific topics (we generate 20 topics for this chapter) from the domain concepts and the topic distributions are used to retrieve the images which are the most relevant in the dataset. Then, we compute the cosine similarity between BoVW to find the nearest images in the retrieved images. We retrieve 100 images generated by the cosine similarity measure and show five images among the images.

7.4.3 Experiment results

We applied the first experiment on 49 concepts. Figure 7.5 and Figure 7.6 shows that the results of precision for 49 concepts with one domain. As we mentioned in the previous section, the purpose of the first experiment is to find the effectiveness of the domains. In the figures, we can see the precision with the one domain search method is higher than the precision with the keywords search. This is because that the one domain even narrows the meaning of tags increasing the precision. The results indicate that the retrieval with one domain can contribute to the precision results. The first experiment shows us to a way to experiment the second experiment assuming that the domains affect the effectiveness of the keyword search results.

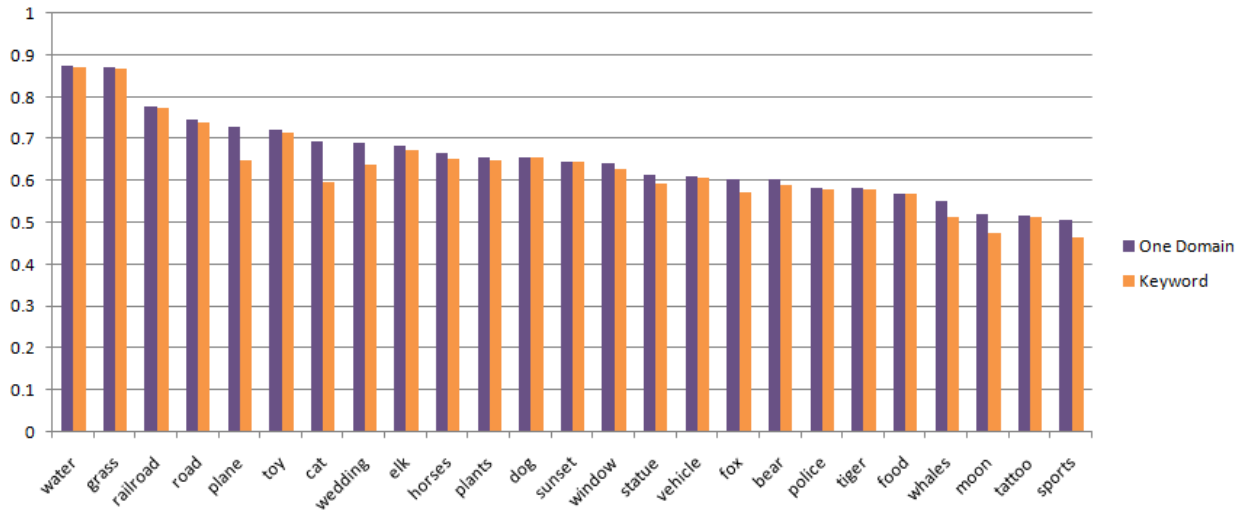

Figure 7.5Precision results of 25 concepts

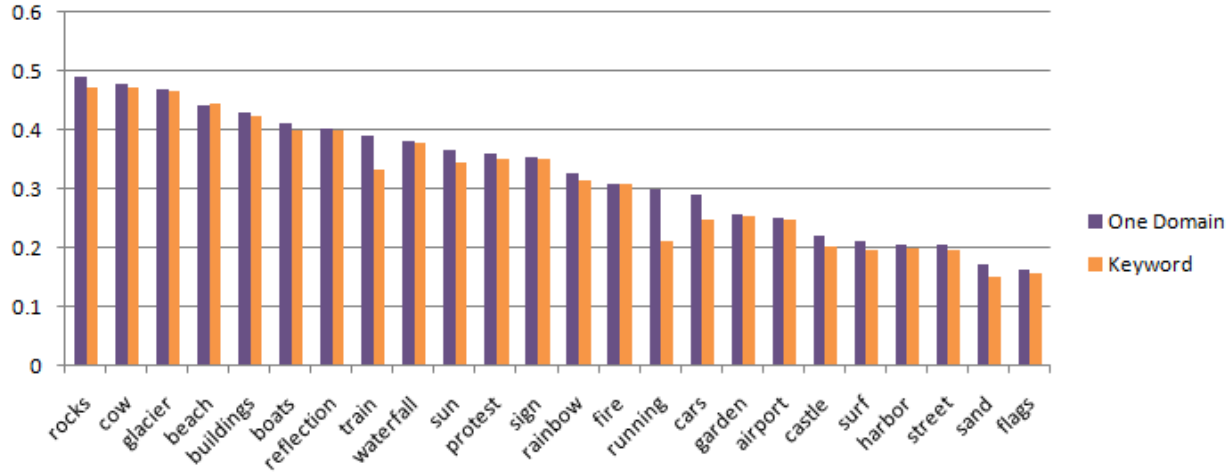

Figure 7.6Precision results of 24 concepts

Figure 7.7 shows experiment results of three different domains (UNDERCARRIAGE, AVIATION, and PERSON) based on NDCG@K [82]. As we already described in the previous section, the second experiment were performed on multi-domains (two domains for this chapter). The experiment results indicate that the proposed process shows the effectiveness on NDCG@K evaluation. The evaluation shows that the results with three different domains have very high NDCG@K values. This is because that we added image features (BoVW) and a topic model (LDA) with two domains (“TRANSPORT” and “UNDERCARRIAGE” or “TRANSPORT” and “AVIATION” or “TRANSPORT” and “PERSON”). The experiment results are very meaningful

because the second experiment shows the effectiveness of the combination of images and tags. Therefore, we can enhance the performance of the tag based social image search when we narrow the meaning of tags with the meaning of images.

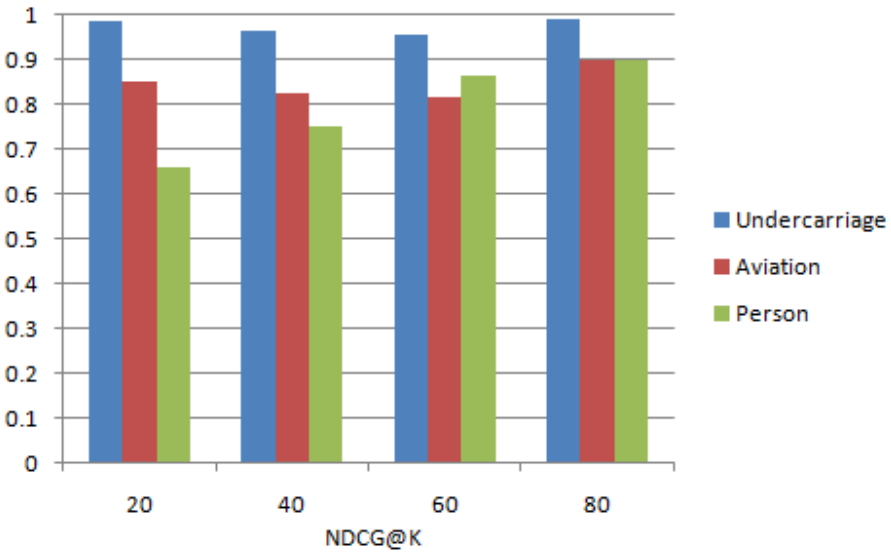

Figure 7.7Experiment results of three domains using NDCG@K

Figure 7.8Examples of the results of the combination of tags and domains

Table 7.3Topics with Tag-Domain pairs

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
night|timeperiod california|california light|physics longexposur|none usa|geography sanfrancisco|sanfrancisco citi|administration canon|religion urban|geography bridg|electronics	girl|person portrait|painting dog|animalscat|animals cute|none babi|person selfportrait|selfportrait kid|person famili|person child|person	uk|geography england|stuartperiod(england) ship|nautical boat|nautical water|geography canada|geography london|london scotland|scotland reflect|physics river|geography	anim|animals natur|psychologicalfeatures wildlif|wildlife bird|animals specanim|none zoo|tourism animalkingdomelit|none impressedbeauti|none naturesfinest|none bear|animals	polit|politics soldier|military protest|sociology polic|military war|sociology usa|geography gun|military armi|military militari|military weapon|military
Topic 6	Topic7	Topic 8	Topic 9	Topic 10
sea|geography beach|geology water|geography ocean|geography sunset|meteorology cloud|meteorology sky|astronomy blue|color island|geography sand|geology	airplan|transport aircraft|transport airport|transport fly|aviation california|california fire|flame aviat|aviation helicopt|transport netherland|geography holland|geography	build|buildings architectur|architecture window|buildings abandon|abandon church|religion old|timeperiod citi|administration architectur|buildings hous|buildings door|buildings	flower|plants macro|computerscience natur|psychologicalfeatures green|color yellow|color garden|buildings closeup|photography red|color spring|timeperiod plant|plants	food|food tabl|furniture shop|buildings red|color chair|furniture kitchen|buildings telephon|telecommunication veget|gastronomy pakistan|geography market|commerce

Figure 7.8 illustrates the results of the combination of tags and domains. Original tags are initially preprocessed to be applied to the domain specific process. After applying two domain specific algorithms, the original tags are replaced with tags-domains sets for social images. The first tag “california” is matched to “California” which is from English Wikipedia entities since WordNet Domains does not include the tag. The WordNet Domains defines a word “male” as “animals”, “factotum”, or “geography”. Since a tag “lion” has “animals” as its domain, “male” is defined as “animals”. Likewise “bigcat” is defined as “Big_cat” and “sandiego” is defined as “San_Diego_California” from English Wikipedia entities. The top third image indicates computers and we can see “apple” is defined as not “PLANTS” but “Apple_inc”. This is because there is a tag “computers” with a domain “COMPUTER_SCIENCE” on the bottom of the tags. Because there are many limitations to retrieve social images by using original tags alone, we generated specific domain concepts for each tag by using proposed process. The generated tag-domain pairs are used to compute topic distributions.

Table 7.3 shows 10 topics generated from LDA model with tag-domain pairs. The tag-domain pairs in each topic represent top 10 θ values generated by 1000 iterations. In general, every topic ideally aims to one concept with tag-domain pairs. Our experiment also assumed that the topics have one concept. The first line of tag-domain pairs in each topic represents the highest value of θ and the last one represents the lowest value of θ among 10 tag-domain pairs. From the tag-domain pairs in the topics we are intuitively able to notice that the pairs are related to each other. For example, in topic 7, the tags (airplan, aircraft, airport, fly, california, fire, aviat, helicopt, netherland, and holand) are considered to related to each other. In some case, however, we may not agree with a tag-domain tag. For example, the second line in topic 9, a tag "macro" matches to a domain "computerscience" rather than something that indicates a large or a whole part. Because "macro" is not a noun or a verb but a prefix our process does not catch about this issue throughout this chapter.

Figure 7.9 shows the five nearest images retrieved by the BoVW with or without domain concepts. Traditionally, it is very difficult to search a similar image with an original image using the BoVW alone because the BoVW only concerns about the features of the image rather than keywords or tags. We apply the BoVW into our DSSP step to retrieve the most similar social images concerning both the image features and the tags. Figure 7.9 (a) shows the five nearest images without domain concepts and the numbers of the bottom of the images represents the cosign similarities between the original images and the retrieved images. Intuitively, we can notify that the image with smaller similarity value on Figure 7.9 (b) or Figure 7.9 (c) is closer to the original image than the image with the highest cosign similarity value on Figure 7.9 (a). This is because we perform the DSSP steps for the second experiment. In other words, for the images in Figure 7.9 (a), we retrieved the nearest images among 269,648 images (we did not retrieve

unknown images when the image links are broken) and, the images in Figure 7.9 (b) and Figure 7.9 (c) are the results among 2,025 and 351 respectively. Figure 7.9 (b) indicates the five nearest images with one domain, "airport|TRANSPORT", and Figure 7.9 (c) indicates the five nearest images with two domains, "man|PERSON" and "airport|TRANSPORT". The results show that the retrieved images with two domains are closer to the original image.

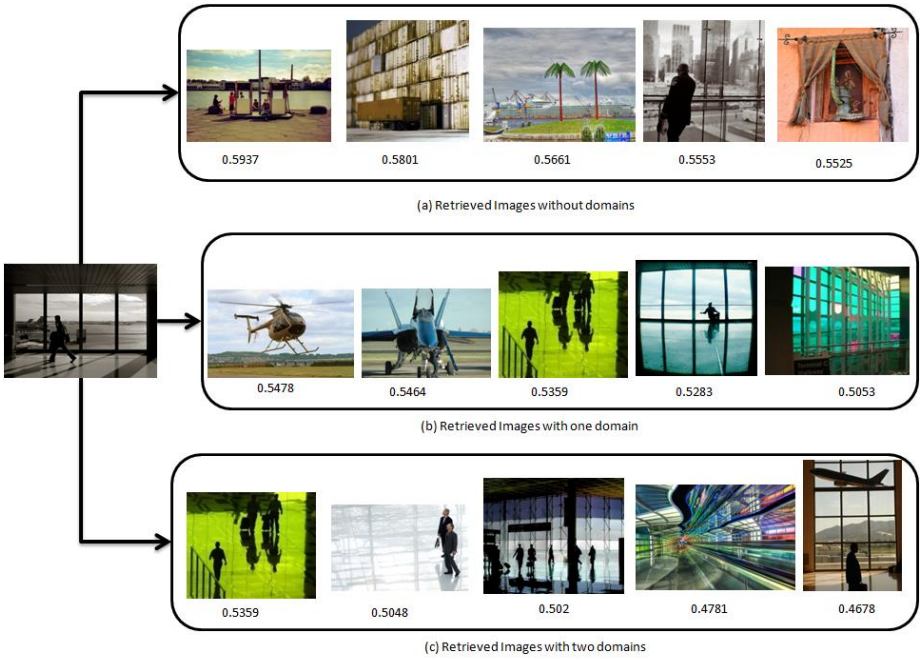

Figure 7.9Top Five nearest images (Domains: Transport and Person)

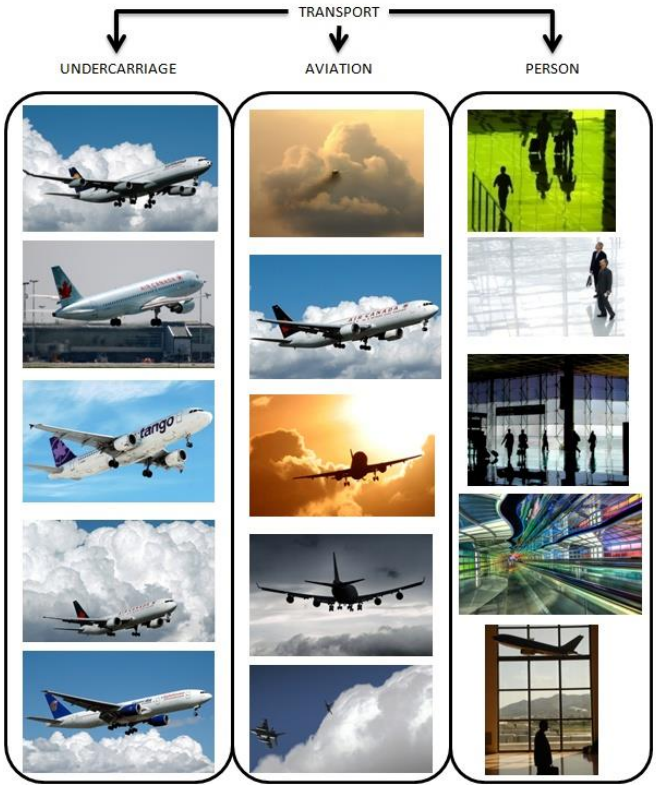

Figure 7.10Five nearest images for UNDERCARRIAGE, AVIATION, and PERSON

Figure 7.10 shows the top five retrieved images for three different specific domain concepts. The first domain concept“TRANSPORT” is combined with the second domain concepts “UNDERCARRAGE”, “AVIATION”, and “PERSON” and then the BoVW computes the cosign similarity between the original images and the retrieved images from the DSSP step. The retrieved images indicate that the DSSP can be used to identify the meaning of images based on the domain concepts and the meaning of the images is more specialized with the two domain concepts.

7.5 Summary

We proposed a domain specific semantic process for large scale images retrieval. Two domain specific algorithms are used to identify the meaning of tags in the images and a topic

model is applied to the results of the tags generating topics for domain concepts. Moreover, the refined tags are combined with the BoVW to reduce the gap between tags and image features. The experimental results showed that the tag-based image retrieval with one domain increases the precision. Also, the tag-based image retrieval with multi-domain shows a high performance on NDCG measure.

We believe that the proposed method is applicable to several areas. For example, recently H. Xie et al. [125] proposed a contextual query expansion model using a visual pattern between two images, which increases retrieval performance even on a large scale database. C. Kang et al. [126] presented a cross-modal matching method for both image and text retrieval and showed the effectiveness of their model.

8 CONCLUSIONS AND FUTURE WORK

8.1 Conclusions

We proposed a domain specific topic model to solve the problems in the area of Information Retrieval in general and LDA model in particular. These problems arise from the difficulty of LDA to specify domain relations and associate them to relevant domains. Two domain specific algorithms are presentedfor handling the domain association difficulty. The proposed algorithms not only narrow semantic concepts down from broad domain knowledge but alsoresolve the unknown domain problem. In order to demonstrate the effectiveness of the proposed model, we conducted various experiments on three different techniques: medical document retrieval, text summarization, and tag-based information retrieval.

Initially, we introduced a medical document retrieval method as a direct application for our domain specific topic model. This method is capable of handling medical specific domains. To verify the effectiveness of the proposed model, we conducted two experiments, document

retrieval and classification. The experimental validation shows that the proposed model outperforms existing models.

We also proposed a novel automatic summarization method that uses our domain specific topic model based on several parameters that include; posting time of tweets, word meanings and four unique characteristics of tweets consisting ofdelivery weights, hashtag relatedness, tf-idf and length of tweets. The experimental results show that the proposed method significantly improves the performance of summarizationoutperforming traditional summarization methods on all aspects.

Finally, we presented a domain specific semantic process for tag-based images retrieval that implements our domain specific topic model. Two domain specific algorithms are appliedfor identifying tag-meanings and generating hidden topics from tags in images. The experiments were conducted to examine the effectiveness of the proposed model and the results showed that the image retrieval combined with our model increases the precision of retrieval measure.

8.2 Future Work

As advances in technology are spreading among endless number of platforms, the convergence of various techniques is necessary to handle the various platforms and maximize their usefulness. In the area of Healthcare Analytics, for example, a paradigm of patient-doctor communication is being shifted to patient-provider communication that keeps ongoing management sources such as daily health checkers, mobile alarms, and cloud record storages. Recently, Seth Earley [100] stated that understanding context is a key portion of Healthcare Analytics providing correct information and mechanisms for patients. We plan to extend our model to the Healthcare Analytics such that the context will be determined by identifying word meanings from unstructured content.

REFERENCES

[1] V. Bush, “As We May Think,” *The Atlantic Monthly*, vol. 176, no. 1, pp. 101-108, July 1945.

[2] C.Cleverdon and M. Keen,“Factors Determining the Performance of Indexing Systems,”*The College of Aeronautics*,vol. 2, Cranfield, England, 1966.

[3] G. Salton, A. Wong, and C. S. Yang, "A Vector Space Model for Automatic Indexing," *Communications of the ACM*, vol. 18, no. 11, pp. 613–620, 1975.

[4] S. Deerwester, S. Dumais, T. Landauer, G. Furnas, and R. Harshman, “Indexing by latent semantic analysis,” *Journal of the American Society of Information Science*, vol. 41, no. 6, pp. 391–407, 1990.

[5] T. Hofmann, “Probabilistic latent semantic indexing,” *Proc. the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR 99*, pp. 50-57, New York, NY, USA, 1999.

[6] D. M. Blei, Y. N. Andrew, and I. J. Michael "Latent dirichlet allocation." *the Journal of machine Learning research*, vol. 3 pp. 993-1022, 2003.

[7] X.R. Wang, A. McCallum, X. Wei, “Topic N-grams: Phrase and Topic Discovery, with an Application to Information Retrieval,” *Proc.IEEE 7th ICDM*, pp. 697-702, 2007.

[8] X. Wei and W.B. Croft, “LDA-Based Document Models for Ad-hoc Retrieval,” *Proc. 29th SIGIR*, pp. 178-185, 2006.

[9] G. Alfio, M. Bernardo, and S. Carlo, "Unsupervised Domain Relevance Estimation for Word Sense Disambiguation", *Proc. the 2004 Conference on Empirical Methods in Natural Language Processing*, Barcelona, Spain, 25-26 July 2004.

[10] B. Magnini, C. Strapparava, G. Pezzulo, and A. Gliozzo, “Using Domain Information for Word Sense Disambiguation,” *in Association for Computational Linguistics SIGLEX Workshop Toulouse*, France, pp. 111-114, 2001.

[11] M. Sahlgren, “An introduction to random indexing,” *Proc. the Methods and Applicatons of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering*. 2005.

[12] G. A. Miller, R. Beckwith, C. D. Fellbaum, D. Gross, and K. Miller, “WordNet: An online lexical database,” *Int. J. Lexicograph*, vo. 3, no. 4, pp. 235–244, 1990.

[13] O. Vechtomova and Y. Wang, “A study of the effect of term proximity on query expansion,”

Journal of Information Science vol. 32, pp. 324–333, Aug, 2006.

[14]E. Voorhees, "Using WordNet to disambiguate Word Senses for Text retrieval,"*ACMSIGIR*, Pittsbourgh, PA, 1993.

[15]J. Gonzalo, F.Verdejo, I. Chugur,and J.Cigarran,"Indexing with WordNet synsetscan improve text retrieval," *Proc.Annual Meeting of the Association for Computational Linguistics and International Conference on Computational Linguistics (COLING/ACL '98) Workshop on Usage of WordNet for Natural Language Processing*, 1998.

[16]C.Carpineto, G. Romano, and V.Giannini, "Improving retrieval feedbackwith multiple term-ranking function combination," *ACM Transactions on Information Systems (TOIS)*vol. 20, no. 3, pp. 259-290, 2002.

[17]H. Cui,J. Wen,J. Nie,and W. Ma, "Probabilistic Query expansion using query logs," *Proc. 11th Internatinal Conference on World Wide Web*, *ACM*, pp. 325-332, Honolulu, Hawaii, 2002.

[18]E. M. Vorhees, "Query expansion using lexical-semantic relations," In: W. Bruce Croft and C. J. van Rijsbergen (Eds.). *Proc. the 17th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '94). Springer*, pp. 61–69, 1994.

[19]P. Kanerva, J. Kristoferson, and A. Holst, "Random Indexing of Text Samples for Latent Semantic Analysis," *Proc. the 22nd Annual Conference of the Cognitive Science Society,* pp. 1036. Mahwah, New Jersey: Erlbaum, 2000.

[20]P. Xie, andE. P. Xing, "Integrating document clustering and topic modeling,"*Proc. the 20th conference on uncertainty in artificial intelligence,* pp. 694–703, 2013.

[21]S. Brody and M. Lapata, "Bayesian word sense induction," *Proc. the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pp. 103–111, 2009.

[22]J. Boyd-Graber, D. Blei, and X. Zhu, "A topic model for word sense disambiguation," *Proc. the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 1024–1033, 2007.

[23]J. Cai, W. S. Lee, and Y. W. The, "Improving word sense disambiguation using topic features," *Proc. the 2007 Joint Conference on EmpiricaMethods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 1015–1023, 2007.

[24]L. Li, B. Roth, and C. Sporleder, "Topic Models for Word Sense Disambiguation and

Token-Based Idiom Detection," *ACL*, 2010.

[25]C. Chemudugunta, A. Holloway, P. Smyth, and M. Steyvers, "Modeling documents by combining semantic concepts with unsupervised statistical learning," *In International Semantic Web Conference*, 2008.

[26]W. Guo and M. Diab, "Semantic topic models: Combining word distributional statistics and dictionary definitions," Proc. the Conference on EmpiricalMethods in Natural Language Processing, EMNLP 2011, pp. 552–561. 2011.

[27]S. C. Wang, and Y. Tanaka, "Topic-oriented query expansion for web search," *in Les Carr; David De Roure; ArunIyengar; Carole A. Goble & Michael Dahlin, ed., WWW , ACM*, pp. 1029-1030, 2006.

[28]Q. T. Zeng, D. Redd, T. Rindflesch, and J. Nebeker, "Synonym, Topic Model and Predicate-Based Query Expansion for Retrieving Clinical Documents," *Proc. AMIA AnnuSymp*, pp. 1050–1059, 2012.

[29]M. Hazewinkel, "Dirichlet distribution," *Encyclopedia of Mathematics, Springer,* 2001.

[30]A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Statist. Soc. B*, vol. 39, pp 1-38, 1997.

[31]A. Bookstein, and D. R. Swanson, "Probabilistic models for automatic indexing," *Journal of the American Society for Information Science*, vol. 26, no. 1, pp. 45-50, 1975.

[32]M. E. Maron, and J. L. Kuhns, "On relevance, probabilistic indexing and information retrieval," *Journal of the ACM*, vol. 7, no. 3, pp, 216-243, 1960.

[33]S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Index by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, 1990.

[34]P. W. Foltz and S. T. Dumais, "An analysis of information filtering methods. Communications of the ACM," vol. 35, no. 12, pp. 51-60, 1992.

[35]J.R. Bellegarda, "Exploiting both local and global constraints for multi-span statistical language modeling," *Proc. ICASSP'98*, vol. 2, pp. 677-680, 1998.

[36]J. Boyd-Graber and D. Blei, "PUTOP: turning predominant senses into a topic model for word sense disambiguation," *Proc. the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pp. 277–281, 2007.

[37]R. Mihalcea, "Large vocabulary unsupervised word sense disambiguation with graph-based

algorithms for sequence data labeling," Proc. the Joint Human Language Technology and Empirical Methods in Natural Language Processing Conference, pp. 411–418, 2005.

[38]D. McCarthy, R. Koeling, J. Weeds, and J. Carroll, "Finding predominant word senses in untagged text," *Proc. the 42nd Meeting of the Association for Computational Linguistics (ACL'04)*, pp. 279–286, 2004.

[39] X. Yao and B. V. Durme, "Nonparametric bayesian word sense induction," *Proc. TextGraphs-6: Graph-based Methods for Natural Language Processing*, pp. 10–14, 2011.

[40]C. Chemudugunta, P. Smyth, and M. Steyvers, "Combining concept hierarchiesand statistical topic models," *Proc. the 17th ACM conference on Information and knowledge management*, pp. 1469–1470, 2008.

[41] J. Lyons, "Semantics 1 & 2," *Cambridge University Press (CUP)*, London and New York, 1977.

[42]H. James and B. Heasley, "Semantics: A Coursebook," *Cambridge University Press (CUP)*, London and New York, 1983.

[43]R. S. Jackendoff, "Semantics and Cognition," *MIT Press*, Cambridge, MA, 1985.

[44]E. Tulving, "Episodic and semantic memory," *In E. Tulving and W. Donaldson (Eds.), Organization of Memory New York: Academic Press*, pp. 381-402, 1972.

[45]J. F. Sowa, "Semantic networks," *Encyclopedia of Artificial Intelligence*, edited by S. C.Shapiro, Wiley, New York, 1987; revised and extended for the second edition, 1992.

[46]Floyd, Robert W. "Assigning Meaning to Programs," In Schwartz, J.T. Mathematical Aspects of Computer Science. Proc. Symposium on Applied Mathematics 19. American Mathematical Society. pp. 19–32. 1967. ISBN 0821867288

[47]R. Arora and B.Ravindran, "Latent Dirichlet Allocation Based Multi-Document Summarization", *Proc. the second workshop on Analytics for noisy unstructured text data (AND)*, pp. 91-97, 2008.

[48]Y. L. Chang and J. T. Chien. "Latent Dirichlet learning for document summarization." *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009.

[49]Y. Lu, and Z. Chengxiang, "Opinion integration through semi-supervised topic modeling." *Proc. the 17th international conference on World Wide Web*. ACM, 2008.

[50]X. Weiand W. B. Croft. "LDA-based document models for ad-hoc retrieval." *Proc. the 29th*

annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2006.

[51]D. Andrzejewski and D.Buttler. "Latent topic feedback for information retrieval." *Proc. the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2011.

[52]L.Bentivogli, P.Forner, B.Magnini, and E.Pianta. "Revising WordNet Domains Hierarchy: Semantics, Coverage, and Balancing", in *COLING 2004 Workshop on "Multilingual Linguistic Resources"*, pp. 101-108, 2004.

[53] J.S. Mitchell, J. Beall, W.E. Matthews, and G.R.New (eds). 1996. Dewey Decimal ClassificationEdition 21 (DDC 21). Forest Press, Albany, NewYork.

[54]Cohen, Aaron M., and William R. Hersh. "A survey of current work in biomedical text mining." *Briefings in bioinformatics* 6, no. 1 (2005): 57-71.

[55]Zhou, Xuezhong, Yonghong Peng, and Baoyan Liu. "Text mining for traditional Chinese medical knowledge discovery: a survey." *Journal of biomedical informatics* 43, no. 4 (2010): 650-660.

[56]Mack, Robert, and Michael Hehenberger. "Text-based knowledge discovery: search and mining of life-sciences documents." *Drug discovery today* 7, no. 11 (2002): S89-S98.

[57]Uramoto, Naohiko, Hirofumi Matsuzawa, Tohru Nagano, Akiko Murakami, Hironori Takeuchi, and Koichi Takeda. "A text-mining system for knowledge discovery from biomedical documents." *IBM Systems Journal* 43, no. 3 (2004): 516-533.

[58]Kim, J-D., Tomoko Ohta, Yuka Tateisi, and Jun'ichi Tsujii. "GENIA corpus—a semantically annotated corpus for bio-textmining." *Bioinformatics* 19, no. suppl 1 (2003): i180-i182.

[59]Shatkay, Hagit, Fengxia Pan, Andrey Rzhetsky, and W. John Wilbur. "Multi-dimensional classification of biomedical text: Toward automated, practical provision of high-utility text to diverse users." *Bioinformatics* 24, no. 18 (2008): 2086-2093.

[60]Huh, Jina, Meliha Yetisgen-Yildiz, and Wanda Pratt. "Text classification for assisting moderators in online health communities." *Journal of biomedical informatics* 46, no. 6 (2013): 998-1005.

[61]Khoo, Christopher SG, Syin Chan, and Yun Niu. "Extracting causal knowledge from a medical database using graphical patterns." In *Proceedings of the 38th Annual Meeting on*

Association for Computational Linguistics, pp. 336-343. Association for Computational Linguistics, 2000.

[62]Holzinger, Andreas, Pinar Yildirim, Michael Geier, and Klaus-Martin Simonic. "Quality-based knowledge discovery from medical text on the web." In *Quality Issues in the Management of Web Information*, pp. 145-158. Springer Berlin Heidelberg, 2013.

[63]Kilgarri, Adam. "Senseval: An exercise in evaluating word sense disambiguation programs." In *Proc. of the first international conference on language resources and evaluation*, pp. 581-588. 1998.

[64]Agirre, Eneko, Oier Lopez de Lacalle, and Aitor Soroa. "Random walks for knowledge-based word sense disambiguation." *Computational Linguistics* 40, no. 1 (2014): 57-84.

[65]Rigau, German, Jordi Atserias, and Eneko Agirre. "Combining unsupervised lexical knowledge methods for word sense disambiguation." In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pp. 48-55. Association for Computational Linguistics, 1997.

[66]Navigli, Roberto. "Meaningful clustering of senses helps boost word sense disambiguation performance." In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pp. 105-112. Association for Computational Linguistics, 2006.

[67]Wang, Quan, Jun Xu, Hang Li, and Nick Craswell. "Regularized latent semantic indexing: A new approach to large-scale topic modeling." *ACM Transactions on Information Systems (TOIS)* 31, no. 1 (2013): 5.

[68]Blei, David M. "Probabilistic topic models." *Communications of the ACM* 55, no. 4 (2012): 77-84

[69]Tan, Ah-Hwee. "Text mining: The state of the art and the challenges." In*Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases*, vol. 8, pp. 65-70. 1999.

[70]Hotho, Andreas, Andreas Nürnberger, and Gerhard Paaß. "A Brief Survey of Text Mining." In *Ldv Forum*, vol. 20, no. 1, pp. 19-62. 2005.

[71]Mei, Qiaozhu, and ChengXiang Zhai. "Discovering evolutionary theme patterns from text: an exploration of temporal text mining." In *Proceedings of the eleventh ACM SIGKDD*

international conference on Knowledge discovery in data mining, pp. 198-207. ACM, 2005.

[72]Baker, L. Douglas, and Andrew Kachites McCallum. "Distributional clustering of words for text classification." In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 96-103. ACM, 1998.

[73]Yarowsky, David. "Unsupervised word sense disambiguation rivaling supervised methods." In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, pp. 189-196. Association for Computational Linguistics, 1995.

[74]Navigli, Roberto. "Word sense disambiguation: A survey." *ACM Computing Surveys (CSUR)* 41, no. 2 (2009): 10.

[75]Miller, George A. "WordNet: a lexical database for English." *Communications of the ACM* 38, no. 11 (1995): 39-41.

[76]Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. "Introduction to wordnet: An on-line lexical database*."*International journal of lexicography* 3, no. 4 (1990): 235-244.

[77]Magnini, Bernardo, and Gabriela Cavaglia. "Integrating Subject Field Codes into WordNet." In *LREC*. 2000.

[78]Lipscomb, Carolyn E. "Medical subject headings (MeSH)." *Bulletin of the Medical Library Association* 88, no. 3 (2000): 265.

[79]Carter-Pokras, Olivia, and Claudia Baquet. "What is a health disparity?."*Public health reports* 117, no. 5 (2002): 426.

[80]Xu, Jinxi, and W. Bruce Croft. "Query expansion using local and global document analysis." In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 4-11. ACM, 1996.

[81]Jones, Karen Sparck. "Automatic keyword classification for information retrieval." (1971). Butterworth, London.

[82]Järvelin, Kalervo, and Jaana Kekäläinen. "Cumulated gain-based evaluation of IR techniques." *ACM Transactions on Information Systems (TOIS)* 20, no. 4 (2002): 422-446.

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

[84]Ron Kohavi: Scaling Up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid. In: Second International Conference on Knoledge Discovery and Data Mining, 202-207, 1996.

[85]George H. John, Pat Langley: Estimating Continuous Distributions in Bayesian Classifiers. In: Eleventh Conference on Uncertainty in Artificial Intelligence, San Mateo, 338-345, 1995.

[86]Chang, Chih-Chung, and Chih-Jen Lin. "LIBSVM: a library for support vector machines." *ACM Transactions on Intelligent Systems and Technology (TIST)*2, no. 3 (2011): 27.

[87]Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. "The WEKA data mining software: an update."*ACM SIGKDD explorations newsletter* 11, no. 1 (2009): 10-18.

[88]B. Sharifi and M. A. Hutton, J. Kalita, "Summarizing microblogs automatically,"in *Proc. HLT/NAACL-10*, pp. 685–688, 2010

[89]D. Inouye, "Multiple post microblog summarization," *Research Final Rep. Colorado Springs, GA: University of Colorado at Colorado Springs*, 2010

[90]D. Radev, H. Jing, M. Sty, D. Tam, "Centroid-based summarization of multiple documents," *Information Processing and Management*, vol. 40, pp. 919–938, 2004

[91]G. Erikan and D. Radev, "LexRank: Graph-based centrality as salience in text summarization," *J. Artif. Intell. Res.*, vol. 22, pp. 457–479, 2004

[92]R. Mihalcea and P. Tarau, "TextRank: Bringing order into texts," in *Proc. EMNLP-04*, pp. 404–411, 2004

[93]D. Inouye and K. K. Jugal, "Comparing twitter summarization algorithms for multiple post summaries," *In Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pp. 298-3069, 2011

[94] L. Vanderwende, H. Suzuki, C. Brockett, and A. Nenkova, "Beyond SumBasic: Task-focused summarization with sentence simplification and lexical expansion," *Information Processing and Management*, vol. 43, no.6, pp. 1606–1618, 2007

[95]R. Zhang, W. Li, D. Gao, Y. Quyang, " Automatic Twitter topic summarization with Speech Acts," *IEEE Trans. on Audio Speech, and Language Processing*, vol. 21, no. 3, pp. 648–658, 2013

[96]S. Lee, S. Shakya, R. Sunderraman, S. Belkasim, "Real Time Micro-Blog Summarization based on Hadoop/HBase," *In Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013 IEEE/WIC/ACM International Joint Conferences on*, vol. 3, pp. 46–49, 2013

[97]S. Ghemawat, H. Gobioff, S. T Leung, “The Google File System,” in *Proc. SOSP 03*, pp. 29–43, 2003

[98]J. Dean and S. Ghemawat, “MapReduce: Simplied Data Processing on Large Clusters,” *Communications of the ACM*, vol. 51, no. 1, pp. 107-113, 2008

[99]F. Chang, J. Dean, S. Ghemawat, W. C. Hsieh, D. A. Wallach, M. Burrows, T. Chandra, A. Fikes, R. E. Gruber, “Bigtable: A Distributed Storage System for Structured Data,” *ACM Trans. Comput. Syst.*, vol. 26, no. 2, pp. 1–26, 2008

[100] Earley, Seth. "The Promise of Healthcare Analytics." *IT Professional* 2 (2015): 7-9

[101] C. Y. Lin, “ROUGE: A Package for Automatic Evaluation of Summaries,”in *Proc. the Workshop on Text Summarization. Branches Out (WAS 2004)*, pp. 74–81, 2004

[102] C. Y. Lin, F. Josef, “Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics,” in *Proc. the 42th Annual Meeting of the Association for Computational Linguistic (ACL 2004)*, pp. 605–612, 2004

[103] A. Sunand S. S. Bhowmick, "Image tag clarity: in search of visual-representative tags for social images." in *Proc.ACM Conf. SIGMM workshop on Social media*, 2009, pp. 19-26.

[104] A. Louis and A. Nenkova, “Automatically evaluating content selection in summarization without human models,” in *Proc. the Empirical Methods in Natural Language Processing*, vol. 1, 2009

[105] N. Rotem, The Open Text Summarizer, http://libots.sourceforge.net, 2003

[106] M. Hassel, “Resource Lean and Portable Automatic Text Summarization,”PhD-Thesis, School of Computer Science and Communication, KTH, ISBN-978-917178-704-0, 2007

[107] L. A. Zadehl, “Fuzzy sets,” in *Information and Control*, vol. 8, no. 3, pp. 338–393, 1965

[108] V. A. Yatsko and T. N. Vishnyakov, “A method for evaluating modern systems of automatic text summarization,” *Automatic Documentation and Mathematical Linguistics*, vol. 41, no. 3, pp. 93–103, 2007

[109] O. Yatsko and B. Smyth, “From social bookmarking to social summarization: an experiment in community-based summary generation,” in *Proc. the 12th international conference on Intelligent user interfaces*, pp. 42–51, 2007

[110] L. H. Reeve, H. Han, A. D. Brooks, “The use of domain-specific concepts in biomedical text summarization,” *Information Processing and Management*, vol. 43, no. 6, pp. 1765–1776, 2007

[111] X. Li, C. G.M. Snoek, M.M. Worring, and A. W.M. Smeulders, "Harvesting social images for bi-concept search," *IEEE Trans. Multimedia,*vol. 14, no. 4, pp. 1091-1104, 2012

[112] Kuo, Yin-Hsi, Wen-Huang Cheng, Hsuan-Tien Lin, and Winston H. Hsu, "Unsupervised semantic feature discovery for image object retrieval and tag refinement," *IEEE Trans. Multimedia,*vol. 14, no. 4, pp. 1079-1090, 2012

[113] Liu, Dong, Shuicheng Yan, Xian-Sheng Hua, and Hong-Jiang Zhang, "Image retagging using collaborative tag propagation," *IEEE Trans. Multimedia,*vol. 13, no. 4, pp. 702-712, 2011

[114] M. Wang, K. Yang, X. S. Hua, and H. J. Zhang, "Towards a relevant and diverse search of social images," *IEEE Trans. Multimedia,*vol.12, no. 8, pp. 829-842, 2010

[115] Gao, Yue, Meng Wang, Zheng-Jun Zha, Jialie Shen, Xuelong Li, and Xindong Wu, "Visual-textual joint relevance learning for tag-based social image search," *IEEE Trans. Image Processing,*vol. 22, no. 1, pp. 363-376, 2013

[116] J. Sang, C. Xu, and J. Liu. "User-aware image tag refinement via ternary semantic analysis," *IEEE Trans. Multimedia,* vol. 14, no. 3, pp. 883-895, 2012

[117] Li, Xirong, Cees GM Snoek, and Marcel Worring, "Learning social tag relevance by neighbor voting," *IEEE Trans. Multimedia,*vol.11, no. 7, pp. 1310-1322, 2009

[118] G. Zhu, S. Yan, and Y. Ma, "Image tag refinement towards low-rank, content-tag prior and error sparsity," in *Proc.ACM Conf. Multimedia,* 2010, pp. 461-470

[119] D. Lu and Q. Li, "Personalized search on Flickr based on searcher'spreference prediction," in Proc. *20th Int Conf. Companion on WorldWide Web (WWW'11)*, 2011

[120] J. Tang, S. Yan, R. Hong, G. J. Qi, and T.S. Chua, "Inferring semantic concepts from community-contributed images and noisy tags," in *Proc. ACM Conf. Multimedia*, 2009, pp. 223-232

[121] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "Nus-wide: A real-world web image database from National University of Singapore,"in *Proc. CIVR*, 2009

[122] Y. Gao, M. Wang, Z. J. Zha, J. Shen, X. Li, and X. Wu, "Visual-textual joint relevance learning for tag-based social image search." *IEEE Trans. Image Processing,*vol. 22, no. 1 pp. 363-376, 2013

[123] V. I. Spitkovskyand A. X. Chang. "A Cross-Lingual Dictionary for English Wikipedia Concepts." In *LREC*, pp. 3168-3175. 2012

[124] Lowe, David G. "Object recognition from local scale-invariant features." In*Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, vol. 2, pp. 1150-1157. Ieee, 1999.

[125] H. Xie, Y. Zhang, J. Tan, L. Guo, and J. Li. "Contextual query expansion for image retrieval." *IEEE Trans.Multimedia,* vol. 16, no. 4. pp. 1104-1114, 2014

[126] C. Kang, S. Xiang, S. Liao, C. Xu, and C. Pan. "Learning Consistent Feature Representation for Cross-Modal Multimedia Retrieval,"*IEEE Trans.Multimedia,* vol. 17, no. 3. pp. 370-381, 2015

[127] Wu, Pengcheng, Steven Chu-Hong Hoi, Peilin Zhao, and Ying He. "Mining social images with distance metric learning for automated image tagging." In*Proceedings of the fourth ACM international conference on Web search and data mining*, pp. 197-206. ACM, 2011.

[128] Liu, Dong, Xian-Sheng Hua, and Hong-Jiang Zhang. "Content-based tag processing for internet social images." *Multimedia Tools and Applications* 51, no. 2 (2011): 723-738.

[129] Liu, Dong, Meng Wang, Linjun Yang, Xian-Sheng Hua, and HongJiang Zhang. "Tag quality improvement for social images." In *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on*, pp. 350-353. IEEE, 2009.

[130] Pantic, Maja, and Alessandro Vinciarelli. "Implicit human-centered tagging [Social Sciences]." *Signal Processing Magazine, IEEE* 26, no. 6 (2009): 173-180.

[131] Huiskes, Mark J., and Michael S. Lew. "The MIR Flickr retrieval evaluation." In*Proceedings of the 1st ACM international conference on Multimedia information retrieval*, pp. 39-43. ACM, 2008.

[132] Sandhaus, Philipp, and Susanne Boll. "Semantic analysis and retrieval in personal and social photo collections." *Multimedia Tools and Applications* 51, no. 1 (2011): 5-33.

[133] Kong, Weihao, Wu-Jun Li, and Minyi Guo. "Manhattan hashing for large-scale image retrieval." In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pp. 45-54. ACM, 2012.

[134] Iskandar, DNF Awang, Jovan Pehcevski, James A. Thom, and Seyed MM Tahaghoghi. "Social media retrieval using image features and structured text." In *Comparative Evaluation of XML Information Retrieval Systems*, pp. 358-372. Springer Berlin Heidelberg, 2007.

[135] S. Lee, S. Belkasim, Y. Zhang. "Multi-document text summarization using topic model and fuzzy logic." In *Machine Learning and Data Mining in Pattern Recognition*, pp. 159-168. Springer Berlin Heidelberg, 2013

[136] J. M. Conroy, J. D. Schlesinger, D. P. O'Leary. "Topic-focused multi-document summarization using an approximate oracle score." In*Proceedings of the COLING/ACL on Main conference poster sessions*, pp. 152-159. Association for Computational Linguistics, 2006

[137] O. Brendan, M. Krieger, D. Ahn. "TweetMotif: Exploratory Search and Topic Summarization for Twitter." In *ICWSM*. 2010

[138] S.Lee, Y. Zhao, M. Masoud, M. Valero, S. Kul, and S. Belkasim. "Domain specific information retrieval and text mining in medical document." In *Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics*, pp. 67-76. ACM, 2015.