JISTaP http://www.jistap.org
Journal of Information Science Theory and Practice
eISSN : 2287-4577 pISSN : 2287-9099

Research Paper
J Inf Sci Theory Pract 5(3): 31-47, 2017
http://dx.doi.org/10.1633/JISTaP.2017.5.3.3

Enhancing the Narrow-down Approach to Large-scale Hierarchical Text Classification with Category Path Information

Heung-Seon Oh
Korea Institute of Science and Technology Information
245 Daehak-ro, Yuseong-gu, Daejeon 305-806, South Korea
E-mail: ohs@kisti.re.kr

Yuchul Jung *
Kumoh National Institute of Technology (KIT)
61 Daehak-ro, Gumi, Gyeongbuk, Korea
E-mail: jyc@kumoh.ac.kr

ABSTRACT
The narrow-down approach, separately composed of search and classification stages, is an effective way of dealing with large-scale hierarchical text classification. Recent approaches introduce methods of incorporating global, local, and path information extracted from web taxonomies in the classification stage. Meanwhile, in the case of utilizing path information, there have been few efforts to address existing limitations and develop more sophisticated methods. In this paper, we propose an expansion method to effectively exploit category path information based on the observation that the existing method is exposed to a term mismatch problem and low discrimination power due to insufficient path information. The key idea of our method is to utilize relevant information not presented on category paths by adding more useful words. We evaluate the effectiveness of our method on state-of-the art narrow-down methods and report the results with in-depth analysis.

Keywords: Hierarchical text classification, Query expansion, Narrow-down approach

Open Access

Accepted date: July 6, 2017
Received date: March 15, 2017

***Corresponding Author:** Yuchul Jung
Assistant Professor
Kumoh National Institute of Technology (KIT)
61 Daehak-ro, Gumi, Gyeongbuk, Korea
E-mail: jyc@kumoh.ac.kr

1. INTRODUCTION

Hierarchical text classification (HTC) aims at classifying documents into a category hierarchy. It is a practical research problem because there are many applications such as online advertising (Broder et al., 2009; Broder, Fontoura, Josifovski, & Riedel, 2007), web search improvement (Zhang et al., 2005), question answering (Cai, Zhou, Liu, & Zhao, 2011; Chan et al., 2013), protein function prediction (Sokolov & Ben-Hur, 2010), and keyword suggestion (Chen, Xue, & Yu, 2008) that rely on the results of HTC to a large scale taxonomy.

In HTC on large-scale web taxonomies, researchers encounter data imbalance and sparseness problems stemming from the internal characteristics of web taxonomies as follows. First, categories spread over a hierarchy from extremely general to specific along its depth. General concepts such as sports and arts appear at the top level while very specific entities such as names of persons and artefacts appear at leaf nodes. Second, two categories with different top-level categories may not be topically distinct because similar topics occur in different paths (i.e., C/D and C'/D' are very similar in c_1=R/A/B/C/D and c_2=R/X/Y/C'/D' even with having different top-level categories). Third, the numbers of documents of categories depend on their popularity on the web. Therefore, there are many categories with a few documents while some categories have many documents.

Traditionally, researchers have focused on developing methods based on machine learning algorithms (Bennett & Nguyen, 2009; Cai & Hofmann, 2004; Gopal & Yang, 2013; Gopal, Yang, & Niculescu-mizil, 2012; Liu et al., 2005; McCallum, Rosenfeld, Mitchell, & Ng, 1998; Sebastiani, 2001; Sun & Lim, 2001; Wang & Lu, 2010; Wang, Zhao, & Lu, 2014). The well-known drawbacks of solely utilizing machine learning are huge computation power and time complexity in order to process large-scale data with a sophisticated algorithm. As a solution, a narrow-down approach (Xue, Xing, Yang, & Yu, 2008) composed of two separate stages, search and classification, was proposed to achieve acceptable levels of effectiveness while increasing efficiency. At the search stage, a small number of candidate categories which are highly relevant to an input document are retained from an entire category hierarchy. At the next stage, classification for final category selection is performed by training a classifier online with documents associated with the candidates selected from the search stage. Based on this idea, narrow-down approach methods are enhanced by incorporating additional information derived from a target hierarchy (Oh, Choi, & Myaeng, 2010, 2011; Oh & Jung, 2014; Oh & Myaeng, 2014). In Oh and Myaeng (2014), three types of information in a hierarchy are defined: local, global, and path information. In category selection, three types of information are employed to find an answer category based on a statistical language modeling framework. Their further work (Oh & Jung, 2014) focused on generating more accurate global information and incorporating local, global, and path information for obtaining a better representation of the input document in a classification aspect (Oh & Myaeng, 2014).

Previously, in a method of using path information, a label language model or label model induced from text of category path was proposed in Oh and Myaeng (2014). It revealed an under-representation phenomenon of label terms (extracted from a category path), which means that the counts of label terms are not as high in documents as expected although they are definitely important in representing categories. The aim of label models is to give more weight to label terms to overcome this situation. In the previous label models, we observed two limitations:

1. First, there exists a term mismatch problem between input documents and label terms. It is one of the well-known problems in information retrieval (IR), since short query terms do not occur in documents (Carpineto & Romano, 2012; Custis & Al-Kofahi, 2007; Zhao & Callan, 2012). In our case, it is the opposite situation where the number of label terms for a category is very small compared with the number of terms for an input document.
2. Second, label models are less discriminative since they have similar probability distributions. This is caused by two reasons. The first reason is that candidates can share many label terms because they are located closely in a hierarchy. Figure 1 shows an example of five candidates retrieved from ODP, a web taxonomy used in our experiments, as an input document and corresponding label models extracted for those candidates. Among 15 unique terms extracted from all candidates, three label terms {sports, winter, skiing} are shared due to a common path *Sports/*

Winter_Sports/Skiing. The second reason is that the maximum likelihood estimation produces similar probability distributions over label terms. Therefore, most of the label terms have zero probabilities while common terms {sports, winter, skiing} have similar probabilities due to low occurrences.

As a novel solution to deal with the term mismatching problem and less discriminative power of the label models, we expand label models by including non-label terms which have strong associations with label terms and estimating probability distributions for label and non-label terms. Our expansion method consists of three steps: translation model construction, non-label term selection, and parameter estimation. We first construct a translation model to capture word-to-word relationships in a category. Then, a set of non-label terms which have strong associations with label terms are selected as expansion terms. Finally, a label model is estimated over label and non-label term sets together. Experiments on the state-of-the art narrow-down methods show the effectiveness of our expansion method in category selection. Our method is built on top of the recent narrow-down approach (Oh & Jung, 2014; Oh & Myaeng, 2014), but it is differentiated with the following contributions:

1. We propose an expansion method for label models to make use of path information more effective. We mainly tackle the term mismatching problem by excavating non-label terms which have a close association with label terms and low discrimination problems by smoothing.
2. We validate the effectiveness of our expansion method by comparing with the state-of-the art narrow-down methods which deal with large-scale web taxonomies using a large data collection, i.e. ODP.

The rest of this paper is organized as follows. In Section 2, key research work on HTC is summarized and briefly compared. In Section 3, we describe our expanded label language models with an introduction of the previous narrow-down approach methods in detail. Section 4 reports the results of the expanded label language models with in-depth analysis by comparing the state-of-the art narrow-down methods. In Section 5, we end this paper with a summary and discussion of future work.

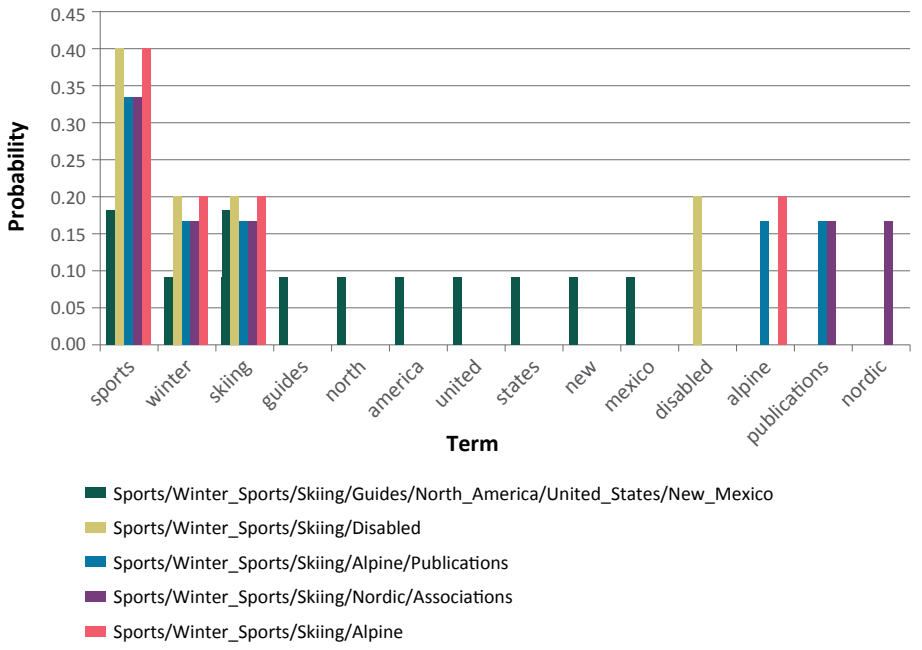

Fig. 1 An example where candidates share a common path *Sports/Winter_Sports/Skiing* and corresponding language models (terms are lowercased and split by under-bar and slash)

2. RELATED WORK

For hierarchical text classification, methods can be categorized into three types of approaches: big-bang, top-down, and narrow-down. In the big-bang approach, a single classifier is trained for all the categories in a hierarchy and an input document is classified into one of them, ignoring the hierarchical structure. Various classification techniques were employed for this approach, including SVMs (Cai & Hofmann, 2004), a centroid classifier (Labrou & Finin, 1999), and a rule-based classifier (Sasaki & Kita, 1998). Koller and Sahami (1997), however, showed that the big-bang approach has difficulty in scaling-up for a web taxonomy in terms of time complexity. A shrinkage method (McCallum et al., 1998) was introduced to deal with the data sparseness problem that may occur with leaf nodes in the big-bang approach. Its main idea is to estimate term probabilities for a leaf node not only based on the documents associated with it but also those associated with its parent nodes up to the root. Mixture weights along the path from a leaf node to the root are calculated using an expectation and maximization algorithm. While the idea was proven to be useful, it has the drawback of huge computation requirements for estimating many parameters. Recent research efforts (Gopal & Yang, 2013; Gopal et al., 2012) proposed methods such as recursively utilizing the dependency between child and parent classes from the root. They incorporate a recursive regularization term into an objective function such as SVMs and logistic regression. Even with this novel idea, it requires a map-reduce framework with many machines.

In the top-down approach, a classifier is trained with the documents associated with each node from the top of a hierarchy. When a new document comes in, it is first classified into one of the top categories directly under the root and then further classified into a node at the next level, which is a child of the node chosen at the previous step. The process is repeated downward along the hierarchy until a stopping condition is met. Several studies adopted this approach with different variations of classification algorithms, such as multiple Bayesian classifiers (Koller & Sahami, 1997) and SVMs (Bennett & Nguyen, 2009; Cai & Hofmann, 2004; Liu et al., 2005; Sun & Lim, 2001).

Liu et al. (2005) compared the big-bang and top-down approaches using SVM on the Yahoo! Directories dataset to show that the top-down approach was more effective and efficient than the big-bang approach. Despite the overall superiority in terms of classification performance, the top-down approach suffers from performance drops at deep levels, caused by errors propagated from higher levels to lower levels. As an effort to deal with the problem, Bennett and Nguyen (2009) devised a method that uses SVMs with the idea of utilizing cross-validation and meta-features. It first performs bottom-up training with cross-validation to produce meta-features that are predictions of lower nodes for each node. When reaching the root, it conducts top-down training with cross-validation to correct document distributions that were fixed according to the hierarchy. This process has the effect of expanding the training data for a node by including misclassified documents at the testing stage through feature vectors consisting of words and meta-features. Even though it achieved remarkable performance improvements on the ODP dataset over the hierarchical SVMs approach (Liu et al., 2005), a drawback is the huge computational overhead required for top-down and bottom-up cross-validations on the entire dataset. More recently, Wang et al. (2014) proposed a meta-top-down approach to large-scale HTC. It is known to be more efficient for reducing the time complexity by considering the top-down training with meta-classifiers. However, their approach is limited to leaf categories.

A narrow-down approach, often referred to as deep classification, was introduced by Xue et al. (2008) to deal with the problems associated with the other two approaches by first cutting down the search space of the entire hierarchy and building a classifier for a small number of resulting categories. The method first employs a search engine to select a set of candidate categories that are highly relevant to an input document to be classified. Trigram language models are constructed for the candidate categories using the documents associated with them for precision-oriented improvements. In order to alleviate the data sparseness problem that occurs with trigrams at deep levels, they proposed the ancestor-assistant strategy. For each candidate, it collects documents not only from the current node but also from those up to the non-shared parent node so that a larger set of documents is used as training data.

The method results in a significant performance improvement, specifically in deeper levels, compared to a hierarchical SVM method on the ODP dataset. Other narrow-down approaches (Oh et al., 2010, 2011) incorporated global information available at the top of the hierarchy and combined it with the local information associated with the candidates for the improvement of classification effectiveness. Oh and Myaeng (2014) proposed passive and aggressive methods by utilizing global information based on a language modeling framework. In addition, a label language model is developed to give weights to label terms in local models by observing those label terms that are not occurring as frequently as expected. Their consecutive research (Oh & Jung, 2014) emphasized that generating accurate global information using ensemble learning is effective. Moreover, it showed a way of incorporating non-local information directly to an input document based on a statistical feedback method by observing that global information has little influence on category selection even with its high accuracy.

3. PROPOSED METHOD

Our method is devised based on the narrow-down approach which consists of candidate search and category selection stages. When an input document comes, a set of relevant candidates are retrieved via the candidate search. Based on the candidates, a final category is selected via category selection. In candidate search, Xue et al. (2008) employed two candidate search strategies, document-based and category-based searches. We chose the category-based search because of better effectiveness (Xue et al., 2008). In the category-based search, a category is presented as a word count vector by concatenating all documents associated with the category. Similarity score against the category with a retrieval model is computed with the word count vector when an input documents comes.

In our experiments, we selected the BM25 weighting model (Robertson & Walker, 1994) to score categories because its effectiveness is already proven in various IR tasks. Based on the candidates, sophisticated classification methods can be employed without much concern for time complexity.

Our key contribution is to devise a new method of using path information in the category selection stage. Prior to introducing our proposed methods, we explain the prerequisite knowledge, statistical language modeling for category selection, and label language models as background.

3.1. Language Models for Category Selection

In IR, statistical language modeling has become a dominant approach to ranking documents (Kurland & Lee, 2006; Lafferty & Zhai, 2001; Ponte & Croft, 1998; Zhai & Lafferty, 2004). The idea of language modeling is to compute the probability of generating a query from a model of a document as in the query likelihood model (Ponte & Croft, 1998) described as follows:

$$score(Q,D) = p(Q|\theta_D) = \prod_{w \in Q} p(w|\theta_D)^{c(w,Q)} \quad (1)$$

where $c(w,Q)$ is a count of term w in query Q.

Another popular ranking function with language models is the KL-divergence scoring method (Lafferty & Zhai, 2001) for which two different language models are derived from a query and a document, respectively, and documents are ranked according to the divergence between the two as follows:

$$score(Q,D) = -KL(\theta_Q||\theta_D) = -\sum_{w} p(w|\theta_Q)\, log \frac{p(w|\theta_Q)}{p(w|\theta_D)} \quad (2)$$

where θ_Q is a query unigram language model.

This scoring function can be used to estimate an approximate probability between two documents for document re-ranking (Kurland & Lee, 2006):

$$score_{KL}(D_1,D_2) = p(D_1|D_2) = exp\,(-KL(\theta_{D_1}||\theta_{D_2})) \quad (3)$$

where θ_{D_1} and θ_{D_2} are unigram document language models. We adopt this scoring function when we compare an input document and a category.

A key challenge in applying language modeling to information retrieval is estimating the probability distributions for a query and a document. A basic method is to compute a maximum likelihood estimate as follows:

$$p_{ML}(w|\theta_D) = \frac{c(w,D)}{|D|} \quad (4)$$

where $c(w,D)$ is a frequency count of a term w in a

document D and $|D|$ is the document length, often measured with the total number of terms in D.

Meanwhile, the problem of the maximum likelihood estimate is assigning a zero probability to unseen words that do not occur in a document. To resolve the limitation, several smoothing methods have been developed to avoid zero probabilities and thus improve retrieval performance. Traditional smoothing methods often use term probabilities in the entire collection in addition to those in a document. The two-stage smoothing method which combines Dirichlet smoothing and Jelink-Mercer is one of the most popular ways to estimate document language models using the entire collection (Zhai & Lafferty, 2004). It is estimated as follows:

$$p_{TS}(w|\theta_D) = (1-\lambda)\cdot\frac{c(w,D)+\mu\cdot p(w|COL)}{\sum_t c(t,D)+\mu} + \lambda\cdot p(w|U) \quad (5)$$

where μ and λ are the Dirichlet prior parameter and the Jelinek-Mercer smoothing parameter, respectively, $c(t,D)$ is a frequency count of a term t in a document D, and COL represents a document collection. The second term $p(w|U)$ is the user's query background language model. When λ=0, two-stage smoothing is the same as Dirichlet smoothing whereas it becomes the same as Jelinek-Mercer smoothing when μ=0. In general, it is approximated by $p(w|COL)$ with insufficient data to estimate $p(w|U)$ even though it is different from $p(w|COL)$.

For category selection, two language models – local model θ_{C_l} and global model θ_{C_g} – are defined for a category C. A local model is derived from the documents associated with the category at hand (a candidate category). A global model is generated from all documents associated with each top-level category, which is a direct child of the root. Note that a category always has local and global models because it must have a path to the root. KL-divergence scoring function is utilized to calculate an approximate probability between an input document and a category.

The goal of category selection is to choose a final category for an input document based on the KL-divergence scoring function:

$$C^{\star} = \underset{C\in H}{argmax}\ \{score(Q,C)\} \quad (6)$$

where Q is an input document and H is the set of candidate categories.

This scoring function is decomposed into two different functions to capture the characteristics of local and global information independently:

$$score(Q,C) = score_{KL}(Q,C_g)\cdot score_{KL}(Q,C_l) \quad (7)$$

$score_{KL}(Q,C_g)$ is a score with θ_{C_g} in a global aspect of a hierarchy while $score_{KL}(Q,C_l)$ is a score with θ_{C_l} in a local aspect. Our focus is how to estimate θ_{C_l} with path information to compute $score_{KL}(Q,C_l)$.

3.2. Label Language Models

The idea of label language models introduced in Oh and Myaeng (2014) is to give more weight to label terms in local models since they are under-represented in associated documents in a hierarchy. Namely, label terms in categories do not occur as frequently in associated documents as we expected although they are definitely important for the purpose of representing categories. Table 1 shows an example of term count and their rank information extracted from associated documents for a category *Sports/Strength_Sports/Bodybuilding/Training* in ODP. We can see that the counts and ranks of label terms are not high unlike our expectation. Due to the under-representation, label terms in a local model have relatively low probabilities.

As a solution to overcome this situation, a local model is defined with a label model as:

Table 1. Count and Rank Information of Label Terms for *Sports/Strength_Sports/Bodybuilding/Training*

Term	sport	strength	bodybuild	train
Count	3	26	49	74
Rank	613	42	14	7

$$p(w|\theta_{C_l}) = (1-\gamma)\cdot\left(\frac{c(w,C_l)+\mu\cdot p(w|COL)}{\sum_t c(t,C_l)+\mu}\right)+\gamma\cdot p_{ML}(w|\theta_{C_{label}}) \quad (8)$$

where p_{ML} $(w|\theta_{C_{label}})$ is a probability of a term w in a label text of a category *C*.

However, label language models over small label terms suffer from term mismatch between an input document and a label model. They may show weak-discriminativeness when the same term counts appear in the text of a category path.

3.3. Expansion Method for Label Language Models

The idea of expanding label models is to include non-label terms occurring in documents which have strong associations with label terms and generate a label model utilizing term counts not only in a category path but also in documents. The generation of our expansion method consists of three steps: translation model construction, non-label term selection, and parameter estimation. First, we should find associations between label and non-label terms. To do that, a translation model p_C $(u|t)$ for a category *C* between a non-label term *u* and a label term *t* is induced using documents associated with category *C*. Several methods can be utilized to build a translation model such as simple co-occurrence between terms (Bai, Song, Bruza, Nie, & Cao, 2005; Schütze & Pedersen, 1997), HAL (hyperspace analogue to language), which is a weighted co-occurrence that is generated by considering a distance between two terms (Bai et al., 2005), mutual information (Karimzadehgan & Zhai, 2010), and a parsimonious translation model (PTM) (Na, Kang, & Lee, 2007). Among them, PTM is adopted to build a non-label term by the label term translation model. PTM stems from a parsimonious document model (PDM) (Hiemstra, Robertson, & Zaragoza, 2004). The goal of PDM is to generate a document model where document-specific terms have high probabilities while collection-specific terms have low probabilities. This is achieved by maximizing the probability of observing terms in a document using an expectation and maximization (EM) algorithm until it converges. A formal estimation of PDM is as follows:

E-step: $$e_w = c(w,D)\cdot\frac{\lambda_{PDM}\cdot p^i(w|\theta_D)}{(1-\lambda)\cdot p(w|\theta_{COL})+\lambda_{PDM}\cdot p^i(w|\theta_D)} \quad (9)$$

M-step: $$p^{i+1}(w|\theta_D) = \frac{e_w}{\sum_t e_t} \quad (10)$$

where λ_{PDM} is a mixture parameter for a document and $p^i(w|\theta_D)$ is a document model in i-th iteration in the EM algorithm.

PTM is an extension of PDM for constructing a translation model. The idea is to generate a translation model over terms for a document which retains a small number of topical terms by automatically discarding non-topical terms. Similarly, we generate a translation model for a category *C* by re-writing p_C $(u|t)$ as follows:

$$p_c(u|t) = \sum_{D\in C} p(u|\theta_D)\cdot p(\theta_D|t) = \sum_{D\in C} p(u|\theta_D)\cdot\frac{p(t|\theta_D)\cdot p(\theta_D)}{p(t)} \quad (11)$$

where $p(t)=\Sigma_{D\in C}p(t|\theta_D)\cdot p(\theta_D)\cdot p_C(u|t)$ can be estimated by collecting $\Sigma_{D\in C}p(u|\theta_D)\cdot p(t|\theta_D)\cdot p(\theta_D)$ and normalizing it. $p(\theta_D)$ is assumed to be a uniform distribution. Translation models can be computed efficiently as we focus on the distributions of non-label terms over few label terms for a category.

In our problem, a non-label term should have strong associations with all label terms to avoid irrelevant information coming from the lack of context in word-to-word relationships. For example, if we construct two translation models for *Sports/Winter_Sports/Skiing/Disabled* and *Sports/Winter_Sports/Skiing/Alpine,* they may be similar to each other because they share a common parent *Sports/Winter_Sports/Skiing*. Specifically, the two models share most label terms except *Alpine* and *Skiing*.

To ensure strong associations with respect to all label terms of interest, a non-label term selection method is devised where non-label term *u* is accepted as an expansion term if $ratio(u)>\tau$. Ratio is defined as follows:

$$Ratio(u) = \frac{\sum_{t\in LT_C} 1 \text{ if } rank(u|t) < R}{|LT_C|} \quad (12)$$

where $rank(u|t)$ is a rank of a non-label term *u* in a translation model for a label term *t, R* is a minimum rank to be considered, and LT_C is a set of label terms extracted from a category *C*.

The intuition behind this selection is that a non-label term should have a certain degree of association with all label terms. The remaining work is to estimate a label model over label and expansion terms. It is obvious that label terms are more important than expansion terms because label terms are selected by humans in con-

structing a hierarchy while expansion terms are selected in an unsupervised way, thus they can have noisy information. Therefore, we generate a mixture of two label models over different term sets:

$$p(w|\theta'_{C_{label}}) = \lambda_{ORG} \cdot p(w|\theta_{C_{label}}) + (1 - \lambda_{ORG}) \cdot p\,(w|\theta_{C^{exp}_{label}}) \quad (13)$$

where λ_{ORG} is a mixture weight for the original label model, $p(w|\theta_{C_{label}})$ is an original label model, and $p(w|\theta_{C^{exp}_{label}})$ is a label model over expanded terms.

To make the models more discriminative, we utilize term counts in documents associated with a category C to estimate $p(w|\theta_{C_{label}})$ and $p(w|\theta_{C^{exp}_{label}})$. As a result, term counts in a category label and corresponding documents are utilized to estimate $p(w|\theta_{C_{label}})$ while add-one smoothing is applied to estimate $p(w|\theta_{C^{exp}_{label}})$ to avoid zero counts of the expansion terms. The parameters for the two label models are estimated as follows:

$$p(w|\theta_{C_{label}}) = \frac{c(w, C_{label}) + c(w, C)}{\sum_t c(t, C_{label}) + c(t, C)} \quad (14)$$

$$p\,(w|\theta_{C^{exp}_{label}}) = \frac{1 + c(w, C)}{\sum_t 1 + c(t, C)} \quad (15)$$

where $c(w,C_{label})$ is a count of a term w in the text of a category C and $c(w,C)$ is a count of a term w in all documents associated with C.

4. EXPERIMENTS

The goal of this paper is to develop a new method which deals with the weakness of the narrow-down approach. Thus, our experiments focus on comparing methods within the narrow-down approach rather than comparing them to big-bang or top-down approaches. To validate the effectiveness of our expansion method, we compare our method with other state-of-the-art narrow-down approaches.

4.1. Data

The Open Directory Project (ODP)[1] dataset was downloaded from the ODP homepage and used for the entire set of experiments. It has a hierarchy of about 70K categories and 4.5M documents associated with the category nodes. At the top level directly connected to the root are 17 categories: *Adult, Arts, Business, Computer, Games, Health, Home, Kids_and_Teens, News, Recreation, Reference, Regional, Science, Shopping, Society, Sports*, and *World*. We went through a filtering process similar to other research (Bennett & Nguyen, 2009; Oh et al., 2011; Xue et al., 2008) to obtain a comparable and meaningful dataset. Documents in the *World* and *Regional* top categories were discarded because they contain non-English pages and geographic distinctions. For the leaf categories whose names are just enumerations of the alphabet such as *A, B,… Z*, we merged them to their parent category because they are topically neither distinct among themselves nor coherent internally. In addition, categories with less than three documents were discarded to ensure that the documents associated with a category are enough for model estimation. Finally, our dataset contains 65,564 categories and 607,944 web pages (documents). A total of 60,000 documents or about 10% of the entire data were selected for testing by following the strategy (Xue et al., 2008) while the rest were used for training. The testing documents were randomly selected proportional to the numbers of the documents in the categories. This is the same collection used in previous work (Oh & Jung, 2014; Oh & Myaeng, 2014). The reason for choosing this test collection is to directly compare our methods to the state-of-the art methods. LSHTC provides several large-scale document collections constructed from ODP and Wikipedia.[2] However, they are not suitable for evaluating our methods because categories and words of documents are encoded to integers. Such encoding is problematic because the idea of our expansion method is based on the use of label terms extracted from category text. Besides, the results are not interpretable.

[1] Open Directory Project, retrieved from http://www.dmoz.org/
[2] Large Scale Hierarchical Text Classification Challenge, retrieved from http://lshtc.iit.demokritos.gr/

Table 2 shows statistics for our dataset. Even though millions of documents exist in ODP, the average number of documents for each category is less than ten as shown in the filtered ODP.

Figures 2 and 3 show the distributions of documents and categories, respectively, over the 15 levels in the filtered ODP. Most documents are spread over from level 3 to level 9. In our experiments, we only report results up to level 9 because they contain about 98% of all the documents.

For the purposes of indexing and retrieving,[3] Terrier, an open source search engine, was employed with stemming and stop-words removal. The BM25 (Robertson & Walker, 1994) was chosen as a retrieval model because its effectiveness is verified in many IR tasks. For category selection, bigrams and trigrams were generated after stemming without stop-words removal. The stemming task is essentially applied because the number of unique n-grams generated would be excessively large.

4.2. Evaluation Measures

Standard class-oriented evaluation is inappropriate for a data set like ODP because the large number of categories makes it very time-consuming and difficult to analyse the results. Therefore, we adopted the level-based evaluation method used in other hierarchical text classification research (Liu et al., 2005; Xue et al., 2008). For example, suppose that a comparison is made between the prediction and answer categories, *Science/Biology/Ecology* and *Science/Biology/Neurobiology/People*, for a given input document. The level-based evaluation matches between the two paths progressively from the top categories (Science on both paths in this case) to the deepest level categories. Whenever the two categories match at a level, it is counted as correct classification. Otherwise it is counted as a mismatch at that level. An example for a partial matching between the two categories is shown in Table 3. The match at each of the first three levels is counted as a correct classification whereas the mismatch at level 4 is counted as a misclassification. This type of matching instances for all the predictions and corresponding answers are accumulated to compute precision and recall at each level.

For evaluation of a classifier, precision, recall, and F1 are often used, where F1 is the harmonic mean of precision and recall:

$$\text{Precision} = \frac{\#\text{ of correct predictions}}{\#\text{ of predictions}}$$

$$\text{Recall} = \frac{\#\text{ of correct predictions}}{\#\text{ of answers}}$$

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Two types of averaging methods have been used with multiple classification instances. For macro-average F1 (MacroF1), F1 scores are averaged for individual answer classes first and then averaged across all the classes. On the other hand, micro-average F1 (MicroF1) is computed using all the individual decisions made for input documents ignoring the answer classes. For a level-based evaluation, MacroF1 of a level is computed by averaging F1 scores for the categories at the level. MicroF1 is computed by collecting decisions of all the documents at the level. To find out about the general tendency across the categories at all the levels in the hierarchy, we employ an additional measure, an overall (OV) score. MacroF1 for OV is computed as follows:

$$OV_{Macro} = \Big(\sum_{l \in Level} \frac{\sum_{c \in C_l} F1(c)}{|C_l|} \Big) / |Level|$$

MicroF1 for OV is identical to the F1 score computed by collecting all decisions in the evaluation and taking an average. Unless mentioned otherwise, performance improvements across different methods reported in this paper are assumed to be based on OV scores.

4.3. Experimental Setting

For the sake of direct comparison with other methods, we chose the same baseline used in the previous work (Oh & Jung, 2014; Oh & Myaeng, 2014). It is the Dirichlet smoothed unigram language model using KL-divergence function with a flat strategy for collect-

[3] Terrier Search Engine, retrieved from http://terrier.org/

Table 2. Data Statistics

	ODP	Filtered ODP (our data set)
Categories	623,319	65,564
Documents	4,538,312	607,944
Levels	20	15
Average # documents per category	7.28	9.27

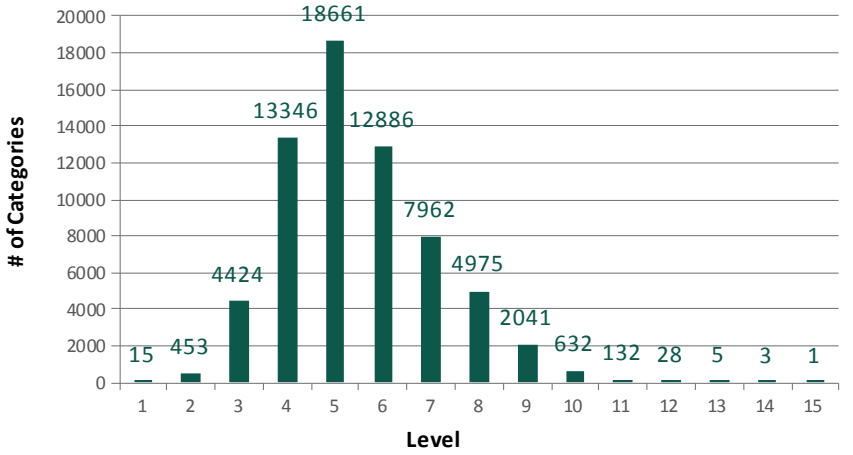

Fig. 2 Category distribution for the filtered ODP

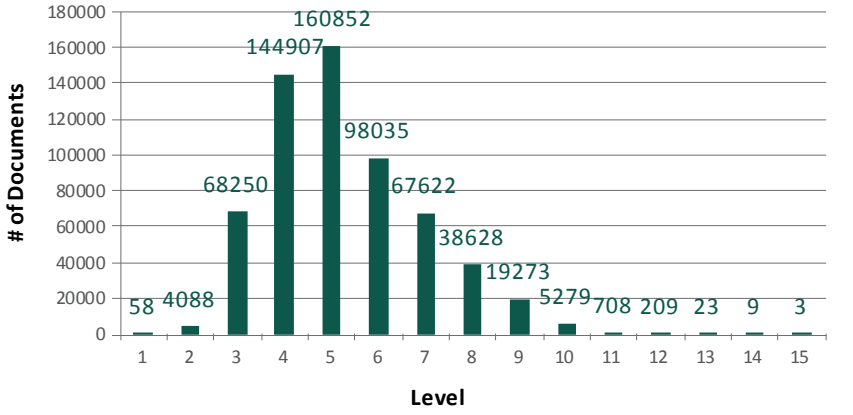

Fig. 3 Document distribution for the filtered ODP

Table 3. Partial Matching between Science/Biology/Ecology and Science/Biology/Neurobiology/People

Level	Partial Prediction	Partial Answer	Correctness Count
1	Science	Science	1
2	Science/Biology	Science/Biology	1
3	Science/Biology/Ecology	Science/Biology/Neurobiology	0
4		Science/Biology/Neurobiology/People	0

ing training data (UKL).

Additionally, we adopt two novel methods which follow the narrow-down approach proposed in Oh and Jung (2014). The first method is a meta-classifier (Meta) with stacking which is a popular ensemble learning framework to combine different algorithms. It can generate accurate global information by combining different top-level classifiers. The second method is query modification modeling (QMM) based on a statistical feedback method. QMM aims at modifying the representation of an input document by incorporating local, global, and path information.

Our designed procedure to compute final score is as follows. First, two scores for an input document Q are obtained in terms of local and global aspects of a category C. Second, Q is updated to Q' using QMM. Note that Q' has a new representation with global, local, and path information. Using Q', the final score for a candidate is computed by combining local and global scores:

1. $score_{Meta}(Q, C) = score_{Meta}(Q, C_g) \cdot score_{KL}(Q, C_l)$
2. $Q' = QMM(Q)$
3. $score_{QMM}(Q, C) = score_{Meta}(Q, C_g) \cdot score_{KL}(Q', C_l)$

Four parameters $\{\gamma,\beta,\gamma_{QMM},K\}$ are considered in Meta and QMM. γ is a control parameter for a label model in a local model shown in equation 8. This term is used to compute $score_{KL}$ (Q,C_l). β is a control parameter for QMM in a new query model. γ_{QMM} is a similar parameter for a label model but used in constructing QMM. K is the number of candidates considered in category selection.

4.4. Results

After a number of experimental runs as in Figure 4, we provide the comparison of the performances using Meta and QMM with the best parameter setting where γ=0.8, β=0.3, and γ_{QMM}=0.1. The performances, both in MicroF1 and MacroF1, are improved over the baseline (UKL) as we increase K from 5 to 25. By increasing the number of candidates in category selection we can expect further improvements.

In our expansion method which is introduced in Section 3.3, four parameters $\{\lambda_{PDM},R,\tau,\lambda_{ORG}\}$ are important factors which can have effects on performances. In constructing translation models, λ_{PDM} is a mixture to estimate PDM using equations 9 and 10. According to the best performance obtained in (Hiemstra et al., 2004), we set λ_{PDM}=0.1. In non-label term selection, two parameters, R and τ, are involved as shown in equation 12. R is a minimum rank of a non-label term to be considered in a translate model. Increasing R indicates that many non-label terms are considered in term selection. τ is a minimum acceptance ratio between 0 and 1. Increasing τ indicates that a non-label term is accepted if it has a strong association with many label terms. We set R=30 and τ=0.5 based on our exhaustive experiments. Finally, a mixture weight, λ_{ORG}, is required for an original label model.

Figure 5 shows the results of varying λ_{ORG} when K is fixed as 5. We can see that expanding label models contributes to the performance improvements, but relying on it too much hurts the performance. The best performance, 0.604 (7.8%) in MicroF1 and 0.368 (13.7%) over UKL, is obtained with λ_{ORG}=0.9.

To check the maximum performance possible, further experiments are conducted with the best performing parameter settings as we increase K. The results show that performances are improved as K becomes large

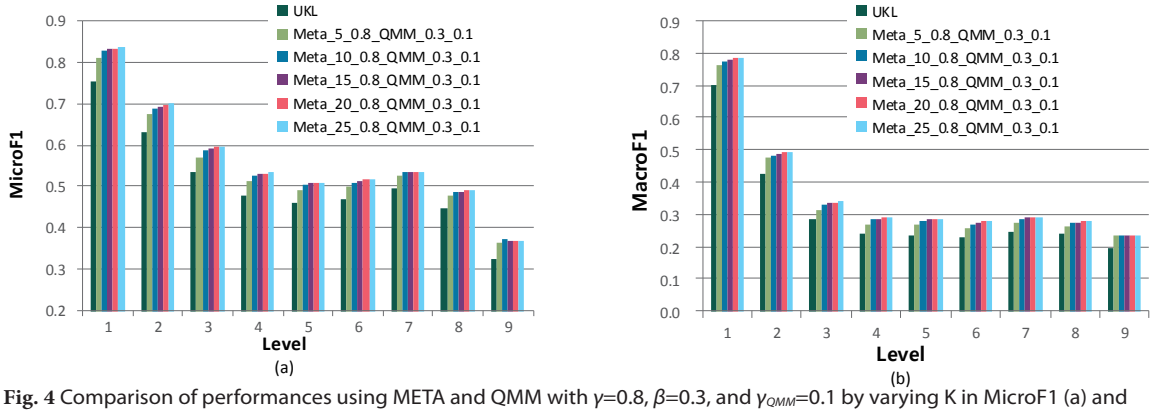

Fig. 4 Comparison of performances using META and QMM with γ=0.8, β=0.3, and γ_{QMM}=0.1 by varying K in MicroF1 (a) and MacroF1 (b)

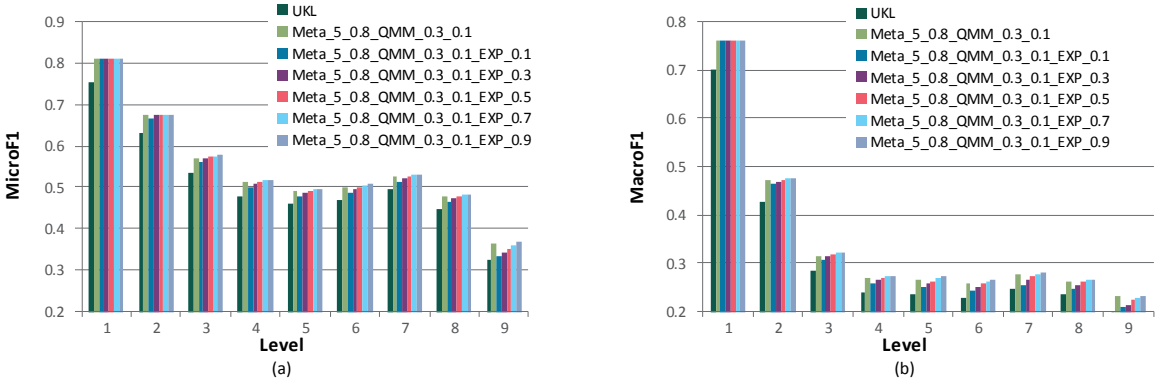

Fig. 5 Comparison of performances using META and QMM with γ=0.8, β=0.3, and γ_{QMM}=0.1 by varying λ_{ORG} with K=5 in MicroF1 (a) and MacroF1 (b)

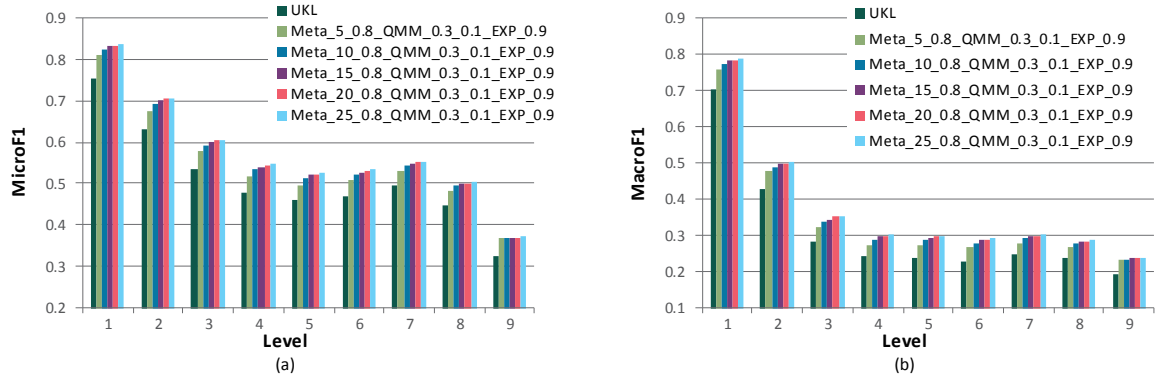

Fig. 6 Comparison among baseline and other variations by varying K in MicroF1 (a) and MacroF1 (b)

as in Figure 6. Its best performances, 0.635 (12.6%) in MicroF1 and 0.395 (20.3%) in MacroF1 over UKL, are obtained with K=25. Meanwhile, increasing K to 50 or 75 or 100 makes little differentiation because the performances are almost stable after K=25.

Although several important parameters are fixed during the experiments, the results were successful in showing the feasibility of expanding path information. We expect that additional improvements can be possible if a modest method which can automatically adjust the parameters is developed.

Table 4 summarizes the best performances of MicroF1 and MacroF1 obtained from the experiments for each of K=5 and K=25. It shows that the proposed method of expanding label models enhances the effectiveness of the state-of-the-art narrow-down approach (i.e., Meta+QMM).

According to the performances where UKL<Meta, we can infer that category selection with global information in conjunction with local and path information works better than local information only. From the performances where Meta<[Meta+QMM], modifying an input document by incorporating global, local, and path information achieves small successes compared with the meta classifier only. However, from [Meta+QMM]<[Meta+QMM+EXP], we can observe that our expansion method makes [Meta+QMM] more robust by including more useful terms. The improvements become larger in both MicroF1 and MacroF1 as K increases.

4.5. In-Depth Analysis of Parameter K

According to Table 4, increasing K contributes to larger improvements. We further analyzed the effectiveness of increasing K in terms of top-level categories. Table 5 shows the performance comparison of K=5 and K=25 with best performing settings at top-level categories. They are listed in descending order with respect to the difference of F1 between K=5 and K=25. The biggest improvement is found in *News* with 20.32% while the smallest one is found in *Adult* with 0.03%. As shown in Figure 7, which compares the differences of F1 measure only, the improvements obtained through the increase of K are more distinct where the categories' F1 measure is less than 0.7 while other categories' improvements are approximately 2-4% except for the *Adult* category.

Based on the observations, we can say that our expansion method with K=25 performs quite well compared to the case of K = 5. We can infer that it assists QMM by adding valid terms selectively regardless of category or subject matter.

5. CONCLUSION

Previous research shows that non-local information such as global and path information play an important role in hierarchical text classification. By observing three limitations of using path information with label language models, term mismatch, and low discrimina-

Table 4. Summary of AVG performances in MicroF1 (above) and MacroF1 (below). Improvements are over the baseline (UKL)

	Baseline (UKL)	**0.564**	
Top-K	Meta	Meta+QMM	Meta+QMM+EXP
5	0.598 (6.0%)	0.604 (7.0%)	0.608 (7.8%)
25	0.622 (10.3%)	0.626 (10.9%)	0.635 (12.6%)

	Baseline (UKL)	**0.328**	
Top-K	Meta	Meta+QMM	Meta+QMM+EXP
5	0.361 (10.2%)	0.368 (12.2%)	0.373 (13.7%)
25	0.385 (17.3%)	0.386 (17.7%)	0.395 (20.3%)

Table 5. Performance Comparison of K=5 and K=25 with the Best Performing Setting at Top-Level Aspect

	K=5			K=25			
Category	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**Imp. in F1**
News	0.5078	0.2500	0.3351	0.6696	0.2885	0.4032	20.32%
Reference	0.6457	0.6290	0.6373	0.6804	0.6836	0.6820	7.01%
Kids_and_Teens	0.4486	0.3870	0.4156	0.5065	0.3932	0.4427	6.52%
Shopping	0.7620	0.7679	0.7649	0.7935	0.7968	0.7952	3.96%
Recreation	0.7837	0.8214	0.8021	0.8240	0.8414	0.8326	3.80%
Science	0.7797	0.7765	0.7781	0.8163	0.7956	0.8058	3.56%
Business	0.8007	0.8217	0.8110	0.8186	0.8595	0.8385	3.39%
Arts	0.8298	0.8361	0.8329	0.8573	0.8597	0.8585	3.07%
Health	0.8361	0.8575	0.8467	0.8564	0.8870	0.8714	2.92%
Home	0.8046	0.8118	0.8082	0.8312	0.8312	0.8312	2.85%
Sports	0.8799	0.9282	0.9034	0.9118	0.9431	0.9272	2.63%
Computers	0.8406	0.7893	0.8142	0.8430	0.8280	0.8354	2.60%
Games	0.8277	0.7953	0.8112	0.8408	0.8210	0.8308	2.42%
Society	0.8653	0.8692	0.8672	0.8895	0.8802	0.8849	2.04%
Adult	0.9435	0.9435	0.9435	0.9395	0.9481	0.9438	0.03%

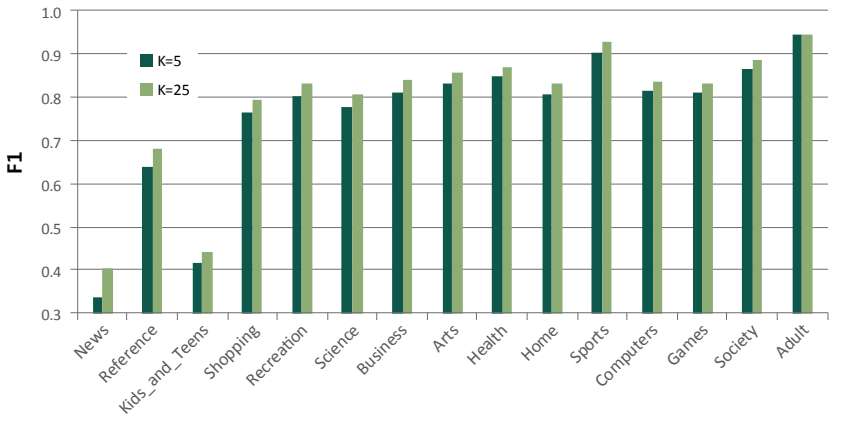

Fig. 7 Performance comparison of K=5 and K=25 (F1 only) over top-level categories

tion power problems of the label language models, we proposed a method to expand label models to overcome the limitations and maximize effectiveness in category selection. Our expansion method is to allow non-label terms which have strong associations with label terms and estimate models over two term sets together. We compare our method based on the most effective narrow-down methods with a large-scale web taxonomy, ODP dataset, used in other research. The best performance, 0.635 (12.6%) in MicroF1 and 0.395 (20.3%), was obtained against the baseline. It outperforms the best performances reported in recent research. It also shows that combining non-local information, i.e. global and category information, with local information is a right choice for dealing with HTC on the narrow-down approach.

Throughout the experiments, the usefulness of appropriately expanding label models is revealed. To improve performance further, we plan to investigate use of the hierarchical structure for label term expansion and use of external collections or taxonomies to make a better representation of an input document.

REFERENCES

Bai, J., Song, D., Bruza, P., Nie, J.-Y., & Cao, G. (2005). Query expansion using term relationships in language models for information retrieval. In *Proceedings of the 14th ACM international conference on Information and knowledge management* (pp. 688-695). New York: ACM. http://doi.org/10.1145/1099554.1099725

Bennett, P. N., & Nguyen, N. (2009). Refined experts: Improving classification in large taxonomies. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval* (pp. 11-18). ACM. Retrieved from http://portal.acm.org/citation.cfm?id=1571946

Broder, A., Ciccolo, P., Gabrilovich, E., Josifovski, V., Metzler, D., Riedel, L., & Yuan, J. (2009). Online expansion of rare queries for sponsored search. In *Proceedings of the 18th international conference on World wide web - WWW '09* (pp. 511-520). New York: ACM Press. http://doi.org/10.1145/1526709.1526778

Broder, A., Fontoura, M., Josifovski, V., & Riedel, L. (2007). A semantic approach to contextual advertising. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '07* (pp. 559-566). New York: ACM Press. http://doi.org/10.1145/1277741.1277837

Cai, L., & Hofmann, T. (2004). Hierarchical document categorization with support vector machines categories and subject descriptors. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management* (pp. 78-87). New York: ACM Press. http://doi.org/10.1145/1031171.1031186

Cai, L., Zhou, G., Liu, K., & Zhao, J. (2011). Large-scale question classification in cQA by leveraging Wikipedia semantic knowledge. In *CIKM'11* (pp. 1321-1330). New York: ACM Press. http://doi.org/10.1145/2063576.2063768

Carpineto, C., & Romano, G. (2012). A survey of automatic query expansion in information retrieval. *ACM Computing Surveys, 44*(1), 1-50. http://doi.org/10.1145/2071389.2071390

Chan, W., Yang, W., Tang, J., Du, J., Zhou, X., & Wang, W. (2013). Community question topic categorization via hierarchical kernelized classification. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management - CIKM '13* (pp. 959-968). New York: ACM Press. http://doi.org/10.1145/2505515.2505676

Chen, Y., Xue, G.-R., & Yu, Y. (2008). Advertising keyword suggestion based on concept hierarchy. In *Proceedings of the international conference on Web search and web data mining - WSDM '08* (pp. 251-260). New York: ACM Press. http://doi.org/10.1145/1341531.1341564

Custis, T., & Al-Kofahi, K. (2007). A new approach for evaluating query expansion. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '07* (pp. 575-582). New York: ACM Press. http://doi.org/10.1145/1277741.1277840

Gopal, S., & Yang, Y. (2013). Recursive regularization for large-scale classification with hierarchical and graphical dependencies. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '13* (pp. 257-265). New York: ACM Press. http://doi.

org/10.1145/2487575.2487644

Gopal, S., Yang, Y., & Niculescu-mizil, A. (2012). Regularization framework for large scale hierarchical classification. In *Large Scale Hierarchical Classification, ECML/PKDD Discovery Challenge Workshop*.

Hiemstra, D., Robertson, S., & Zaragoza, H. (2004). Parsimonious language models for information retrieval. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 178-185). New York: ACM Press. http://doi.org/10.1145/1008992.1009025

Karimzadehgan, M., & Zhai, C. (2010). Estimation of statistical translation models based on mutual information for ad hoc information retrieval. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval - SIGIR '10 (pp. 323-330)*. New York: ACM Press. http://doi.org/10.1145/1835449.1835505

Koller, D., & Sahami, M. (1997). Hierarchically classifying documents using very few words. In *Proceedings of the 4th International Conference on Machine Learning* (pp. 170-178). Morgan Kaufmann Publishers Inc. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.31.2455&rep=rep1&type=pdf

Kurland, O., & Lee, L. (2006). PageRank without hyperlinks: Structural re-ranking using links induced by language models. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '05* (pp. 306-313). New York: ACM Press. http://doi.org/10.1145/1076034.1076087

Labrou, Y., & Finin, T. (1999). Yahoo! as an ontology. In *Proceedings of the eighth international conference on Information and knowledge management - CIKM '99* (pp. 180-187). New York: ACM Press. http://doi.org/10.1145/319950.319976

Lafferty, J., & Zhai, C. (2001). Document language models, query models, and risk minimization for information retrieval. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '01* (pp. 111-119). New York: ACM Press. http://doi.org/10.1145/383952.383970

Liu, T.-Y., Yang, Y., Wan, H., Zeng, H.-J., Chen, Z., & Ma, W.-Y. (2005, June 1). Support vector machines classification with a very large-scale taxonomy. *ACM SIGKDD Explorations Newsletter*. ACM. http://doi.org/10.1145/1089815.1089821

McCallum, A., Rosenfeld, R., Mitchell, T. M., & Ng, A. Y. A. Y. (1998). Improving text classification by shrinkage in a hierarchy of classes. In *Proceedings of the Fifteenth International Conference on Machine Learning* (pp. 359-367). Citeseer. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.73.5412&rep=rep1&type=pdf

Na, S. H., Kang, I. S., & Lee, J. H. (2007). Parsimonious translation models for information retrieval. *Information Processing and Management, 43*(1), 121-145. http://doi.org/10.1016/j.ipm.2006.04.005

Oh, H.-S., Choi, Y., & Myaeng, S.-H. (2010). Combining global and local information for enhanced deep classification. In *Proceedings of the 2010 ACM Symposium on Applied Computing - SAC '10* (pp. 1760-1767). New York: ACM Press. http://doi.org/10.1145/1774088.1774463

Oh, H.-S., Choi, Y., & Myaeng, S.-H. (2011). Text classification for a large-scale taxonomy using dynamically mixed local and global models for a node. In *Proceedings of the 33rd European conference on Advances in information retrieval* (pp. 7-18). Springer. http://doi.org/10.1007/978-3-642-20161-5_4

Oh, H.-S., & Jung, Y. (2014). External methods to address limitations of using global information on the narrow-down approach for hierarchical text classification. *Journal of Information Science, 40*(5), 688-708. http://doi.org/10.1177/0165551514544626

Oh, H.-S., & Myaeng, S.-H. (2014). Utilizing global and path information with language modelling for hierarchical text classification. *Journal of Information Science, 40*(2), 127-145. http://doi.org/10.1177/0165551513507415

Ponte, J. M., & Croft, W. B. (1998). A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '98* (pp. 275-281). New York: ACM Press. http://doi.org/10.1145/290941.291008

Robertson, S., & Walker, S. (1994). Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information*

retrieval (pp. 232-241). New York: Springer-Verlag. Retrieved from http://dl.acm.org/citation.cfm?id=188490.188561

Sasaki, M., & Kita, K. (1998). Rule-based text categorization using hierarchical categories. In *IEEE International Conference on Systems, Man, and Cybernetics* (Vol. 3, pp. 2827-2830). IEEE. http://doi.org/10.1109/ICSMC.1998.725090

Schütze, H., & Pedersen, J. O. (1997). A cooccurrence-based thesaurus and two applications to information retrieval. *Information Processing & Management, 33*(3), 307-318. http://doi.org/10.1016/S0306-4573(96)00068-4

Sebastiani, F. (2001). Machine learning in automated text categorization. *ACM Computing Surveys, 34*(1), 1-47. http://doi.org/10.1145/505282.505283

Sokolov, A., & Ben-Hur, A. (2010). Hierarchical classification of gene ontology terms using the GOstruct method. *Journal of Bioinformatics and Computational Biology, 8*(2), 357-76. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/20401950

Sun, A. S. A., & Lim, E.-P. L. E.-P. (2001). Hierarchical text classification and evaluation. In *Proceedings 2001 IEEE International Conference on Data Mining* (pp. 521-528). IEEE Computer Society. http://doi.org/10.1109/ICDM.2001.989560

Wang, X.-L., & Lu, B.-L. (2010). Flatten hierarchies for large-scale hierarchical text categorization. In *2010 Fifth International Conference on Digital Information Management (ICDIM)* (pp. 139-144). IEEE. http://doi.org/10.1109/ICDIM.2010.5664247

Wang, X. L., Zhao, H., & Lu, B. L. (2014). A meta-top-down method for large-scale hierarchical classification. *IEEE Transactions on Knowledge and Data Engineering, 26*(3), 500-513. http://doi.org/10.1109/TKDE.2013.30

Xue, G. R., Xing, D., Yang, Q., & Yu, Y. (2008). Deep classification in large-scale text hierarchies. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 619-626). New York: ACM Press. http://doi.org/10.1145/1390334.1390440

Zhai, C., & Lafferty, J. (2004). A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems, 22*(2), 179-214. http://doi.org/10.1145/984321.984322

Zhang, B., Li, H., Liu, Y., Ji, L., Xi, W., Fan, W., & Ma, W.-Y. (2005). Improving web search results using affinity graph. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 504-511). New York: ACM Press. http://doi.org/10.1145/1076034.1076120

Zhao, L., & Callan, J. (2012). Automatic term mismatch diagnosis for selective query expansion. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval - SIGIR '12* (pp. 515-524). New York: ACM Press. http://doi.org/10.1145/2348283.2348354