Extension of Language Model to Solve Inconsistency, Incompleteness, and Short Query in the Collection of Cultural Heritage

K. L. Tan, C. K Lim

Department of Computer Science, Faculty of Art, Computing and Creative Industry,Sultan Idris Education University, Tanjong Malim, Perak Darul Ridzuan 35900, Malaysia.

KianLam@fskik.upsi.edu.my

***Abstract*—With the explosive growth of online information such as email messages, news articles, and scientific literature, many institutions and museums are converting their cultural collections from physical data to digital format. However, this conversion results in the issues of inconsistency and incompleteness. Besides, the usage of inaccurate keywords also results in short query problem. Most of the time, the inconsistency and incompleteness are caused by the aggregation fault in annotating a document itself while the short query problem is caused by naive user who has prior knowledge and experience in cultural heritage domain. In this paper, we presented an approach to solve the problem of inconsistency, incompleteness and short query by incorporating the Term Similarity Matrix into the Language Model. Our approach is tested on the Cultural Heritage in CLEF (CHiC) collection, which consists of short queries and documents. The results show that the proposed approach is effective and has improved the accuracy in retrieval time.**

***Index Terms*— Cultural heritage; Information retrieval; Language Model.**

I. Introduction

With the rise of new communication technologies, institutions and museums need to convert all the information of cultural heritage to digital format. It becomes easier for people around the world to access the information of cultural heritage and make this information accessible to the global research community. In order to ease the job of the user to search the cultural heritage from one website to another website, Europeana takes the initiative to collect the metadata, which is a digital format, to represent the cultural heritage across many European Union (EU) member states. Basically, Europeana relies on aggregators who work at the national or domain level to prepare the metadata and to transfer the metadata back to Europeana. Relying on such aggregators indirectly causes the problem of inconsistency and incompleteness in the metadata. For example, if one annotator uses “syriac” to describe an object and another annotator uses “language of ancient syria” to describe the same object, then two annotations are inconsistent with regard to the content of the annotation. Inconsistency may also refer to the structure of the annotations themselves. For instance, some annotators might insert all the information into one description field and others may split it into multiple metadata fields, such as the field of description and the title. In such cases, the information of an object may differ depending on the human annotators. As a result, it is hard for a user to search objects since the characteristics of the cultural heritage are not formatted in the same way.

In this paper, we consider such problems in the context of Information Retrieval (IR). Language Models (LM) for IR has been proven that it is a very effective on text retrieval based on [17]. The extension that we propose in this paper is to integrate term links (Term Similarity Matrix) into the LM based on Dirichlet smoothing that is the most effective Smoothing technique. Our proposal has the following advantages: a) it is easy and simple to generate term links based on statistical information if compared to synthetic queries in [3] or mutual information [10], which considered as heavy method and b) it is a light weight integration in the LM.

II. The Term Intersection Problem

In the past, a number of IR models such as Vector Space Model (VSM) [22, 23], Probabilistic Model [20, 1] and LM [17, 28], which based on term intersection approach, have been proposed. The term intersection is the approach where both the document and query should share the same terms. Although this approach provides a good result in terms of speed and accuracy, it does not solve the problem of term mismatch, in which the document does not compromise the same terms with the query.

For example, a user is searching for the information about a “schlesian map” and submits the query:

q = (schlesian, map)

and IRS considering the documents below:

d1 = (map, germany)
d2 = (china, map)
d3 = (Germany, Silesia)

The Information Retrieval System (IRS) assigns a very similar Retrieval Status Value (RSV) to d1 and d2, which are highly dependent on the indexing weights because these documents contain similar terms as the query, which is the “map”. However, we know that d2 is surely not relevant since d2 contains the information of “china map” and not the information of “schlesien map”. In addition, we can defend that d3 is more relevant than d2, if it is compared to the needs of the user. In the context of IR, this problem is called as the term mismatch, where the term is a mismatch between the query and the document. For example, a user is searching for the information about a “schlesian map” and submits the query.

III. Approaches To The Term Intersection Problem

There are several techniques, such as query expansion [25, 26], relevance feedback [21, 12], dimension reduction [18, 11, 8, 2, 9], statistical translation model [3, 10, 28], and others [4, 27, 6], which are considered as approaches to navigate and to explore the material of cultural heritage domain.

A. Query Expansion

Query expansion (QE) is an approach to reformulate the query to improve the retrieval performance. Basically, QE relies on techniques, such as finding the synonyms of terms through thesaurus fixing the spelling errors from the query and finding another terms and automatically adding these terms to the query [25]. Recent work from Zhao and Callan [26] proposed an automatic diagnosis tool to address the problem of term mismatch. Unfortunately, the automatic diagnosis tool required manual query reformulation instead of automatic query reformulation.

B. Relevance Feedback

The idea of relevance feedback is to involve the user in the IR process in order to improve the final result. Usually, relevance feedback consists of three types such as 1) explicit feedback, 2) implicit feedback and 3) pseudo or blind feedback [13]. Rocchio algorithm [21] is the classic algorithm for implementing explicit feedback, which enables the user to select terms to be added to the original query terms by automatically extracting them from the documents.

Implicit feedback is a method used to incorporate the user behavior process such as duration of time to view a document into a IR process while blind feedback provides a method for automatic local analysis. It automates the manual part of the Rocchio algorithm without an extended interaction with the user. This method performs normal retrieval to find an initial set of relevant documents and makes the assumption that the top k ranked documents are the most relevant.

Lavrenko and Croft [12] proposed an approach to estimate a relevance model with no training data, which used only the query alone. The main problem of implicit and explicit relevance feedback is that it relies on accurate ways of finding term relation in order to avoid the problem query drift.

C. Dimension Reduction

Dimension reduction is the process to reduce the number of random variable that the query and the document refer to the same concept but using different terms. This can be achieved by using thesaurus [9], concept based approach [2], stemming [18,11], and latent semantic indexing [8]. All these techniques proposed different strategies to reduce the chances that the query and document refer to the same concept but using different terms. In the later development, Peng et al.[16] performed stemming according to the context of the query, which helps to improve the accuracy and the performance of retrieval than the query independent stemmers such as Porter[18] and Krovetz [11]. Deerwester et al. [8] proposed to solve the dimension reduction by representing the terms and the documents in a latent semantic space, where the terms that are similar in the space tend to be the terms that not only co-occur in the documents, but also appear in similar contexts.

D. Statistical Translation Model

Statistical Translation Model is a model where all the translation are generated on the basis of statistical models. The idea is based on information theory where a document is translated according to the probability distribution P(u|v), which gives the probability that word, v can be semantically translated to word, u in order to address the problem of term intersection [3,28]. Unfortunately, Statistical Translation Model requires the training data and some relevant query-document pairs where the documents are relevant to the query.

E. Others

Carmel et al. [4] and Yogev et al. [27] proposed an entity oriented search (EoS), which is based on a combination of an expressive query language, faceted search, and the entity relationship (ER) graph navigation. In addition, Clough et al. [6] proposed to model a path or trail, which provides a way for the users to access and to utilize the contents of digital libraries that enrich the experiences of these resources. The main goal of these works is to help the user to navigate and explore the material of cultural heritage domain. The accuracy of the works from [4, 26] is highly dependent on the availability and the quality of the entity extraction tools.

Based on the example in Section 2, all the approaches can help to retrieve d3 if the term “silesia” or “germany” is added into the query. In a nutshell, various techniques have been proposed to solve the problem of term mismatch and all the approaches tend to improve the accuracy of the matching process. There are a number of approaches to solve the term mismatch problem by using LM. The recent works from Berger and Lafferty [3] and Karimzadehgan and Zhai [10] proposed to use statistical translation model to solve the term mismatch problem. The main different between these two works is Berger and Lafferty [3] used synthetic queries, while Karimzadehgan and Zhai [10] used mutual information to generate the relationship between the two terms.

IV. Proposal

As mentioned earlier, our goal is to integrate the Term Similarity Matrix into the LM. After the reviews of Crestani [7], Karimzadehgan and Zhai [10], we considered the problems and proposed to use the approach as shown below:

- We proposed to use the maximum or the highest value instead the total value from the term similarity between the terms from the query with the terms from document. Besides, we only considered the point of view of a query if we cannot find a term in the document, then we consider the closest semantic terms from the document.

- We proposed to use statistical approach rather than probability approach in order to avoid the value of P(w|u) is higher than P(w|w) for a term w obtained by Karimzadehgan and Zhai [10].

Before we build the Term Similarity Matrix, we need to find the links between all the terms in the collection naming V of this vocabulary.

$$t, t' \in V, 0 \leq \mathrm{Sim}(t, t') \leq 1 \qquad (1)$$

1. Sim(t, t′) = 0, there is no link between the term t and t′
2. Sim(t, t′) < 1, there is a link between the term t and t′
3. Sim(t, t′) = 1, there is an exact match between the term t and t′

Basically, we made the first assumption that two terms are considered linked to each other if both terms co-occur in the same context. Such assumption is similar to Peat [15] who assumed that a pair of terms that co-occur frequently in the document is about the same subject. Thus, the data of term co-occurrence obtained from the collection can be used to identify some of the semantic relationships that exist between terms. Based on the previous example, if we have the Term Similarity Matrix, which contains the link between the term of "schlesien"and"silesia", then the IRS will return d1, d2, and d3.

The main idea of this research is to integrate the Term Similarity Matrix into the current Dirichlet formula. Firstly, we need to assume that a term, w is w′ ∈ d can play the role of w where w is w ∈ q during the matching process. More specifically, we consider that if w does not occur in the initial document, but d occurs in the document d_{ext}, which is the result of the extension of d according to the query and some knowledge, the probability of the term w′ is defined according to the extended document d_{ext}. The knowledge assumes to form a symmetrical similarity function, which is Sim : V × V → [0, 1], that denotes the strength of the similarity between two terms from the vocabulary (the larger the value, the higher the strength). We proposed that: ∀w, w′ ∈ V, Sim(w, w′) = 1 if the exact matching between w with w′, and ∀w,w′ ∈ V,Sim(w,w′) = 0 if w does not contain any link with w′.

In order to avoid any complex extensions (see the state of the art), we defined the following constraints:

- One query term, w must only impact occurrences of one document term w′;

To achieve this, we used some simple and sensible heuristics:

1. If a query term, w occurs in a document, d, then the term will not change the length of the document;
2. If a query term, w does not occur in a document d but the term w contains a link with w′ (term from document), then we define as:

$$w''=\arg_{\max w'} \in_{d,w'} \neq_{w} \mathrm{Sim}(w,w') \qquad (2)$$

as the term from the document will serve as the basic count of the pseudo occurrences of w in d as:

$$c(w''; d).\mathrm{Sim}(w'', w) \qquad (3)$$

This pseudo occurrences of the term w″ are then included into the size of the extended document;

3. If a query term, w does not occur in the document and does not contain any link, then it's occurrences is counted in the extended document.

Eventually, using the usual set of notations for the terms that occur in the document and the query, then the new length of the document ($|d_{ext}|$) is:

$$|d_{ext}|= \sum_{w\in d\cap q} c(w; d) + \sum_{w''\in d\backslash q;\mathrm{Sim}(w,w'')\neq 0} c(w''; d).\mathrm{Sim}(w'',w) + \sum_{w'\in d\backslash q;\mathrm{Sim}(w,w')=0} c(w'; d) \qquad (4)$$

with w" defined above for one query term, w so that:

$$w'' =\operatorname{argmax}_{w'\in d,w'\neq w}\mathrm{Sim}(w,w') \qquad (5)$$

Using the fact above, the expression of (|dext|) can be easily simplified into:

$$|dext| = |d| + \sum_{w''\in d\backslash q;\ \mathrm{Sim}(w,\ w'')\ \neq 0} c(w''; d).\mathrm{Sim}(w'', w) \qquad (6)$$

Note that our proposal is to extend the document according to the query. With all the elements described above, the extended Dirichlet Smoothing leads to the following probability for the term, w of the vocabulary V in the document extended dext according to a query q, note that Pμ(w|dext) is defined as:

- if w ∈ d ∩ q:

$$P\mu(w|dext) = \frac{c(w; d) + \mu P(w'|C)}{|dext| + \mu} \qquad (7)$$

- if ∃ w″ ∈ d \ q;Sim(w,w″) ≠ 0:

$$P\mu(w|dext) = \frac{c(w''; d).\mathrm{Sim}(w, w'') + \mu P(w''|C)}{|dext| + \mu} \qquad (8)$$

with $w'' = \mathrm{argma}_{x\ w'\in d,\ w'\neq w}\ \mathrm{Sim}(w,w')$.

- if∃ \w″ ∈ d\q;Sim(w,w″)≠0

$$P\mu(w|dext) = \frac{c(w; d) + \mu P(w|C)}{|dext| + \mu} \qquad (9)$$

with $w'' = \operatorname{argmax}_{w'\in d,\ w'\neq w}\ \mathrm{Sim}(w,w')$.

In a specific case, when all the query terms from q occur in the document, d the first case in the above is used where $|d_{ext}| = |d|$ leads to $P_{\mu}(w|d) = P_{\mu}(w|d_{ext})$.

V. Term Similarity Matrix Based On Statistical Approach

In this section, we propose an easier and lightweight way if compared to Expected Mutual Information Measure (EMIM) [7] to compute the Term Similarity Matrix. Similarity between terms can be represented in a variety ways. In our approach, we used Confidence Coefficient (CC), Tanimoto Similarity (TS), Dice Coefficient (DC), Cosine Similarity (CS) and Overlap Coefficient (OC) to generate the statistical information [19]. The CC between term w_i and w_j are calculated as follows:

$$\mathrm{Sim}_{CC}(w_i, w_j) = \frac{n\,(wi \cap wj)}{n(wi)} \quad (10)$$

where $n(w_i)$ is the number of term (w_i) in the corpus, and $n(w_i \cap w_j)$ is the number of terms that term w_i co-occur together with w_j in the corpus.

The TS between term term w_i and w_j are calculated as follows:

$$\mathrm{Sim}_{TS}(w_i, w_j) = \frac{n\,(wi \cap wj)}{n(wi)+n(wj)-n(wi \cap wj)} \quad (11)$$

The DC between term term w_i and w_j are calculated as follows:

$$\mathrm{Sim}_{DC}(w_i, w_j) = \frac{2n|wi \cap wj|}{n(wi)+n(wj)} \quad (12)$$

The CS between term term w_i and w_j are calculated as follows:

$$\mathrm{Sim}_{CS}(w_i, w_j) = \sqrt{\frac{n\,(wi \cap wj)}{n(wi).n(wj)}} \quad (13)$$

The OC between term term w_i and w_j are calculated as follows:

$$\mathrm{Sim}_{OC}(w_i, w_j) = \frac{n\,(wi \cap wj)}{min(n(wi),n(wj))} \quad (14)$$

Although mutual information is the best to extract semantic similarity information, mutual information is the most expensive in terms of computation (quadratic complexity) according to Maekines et al.[14]. In addition, mutual information computes all the possible combinations of attribute pairs for two given terms are involved while CC, TS, DC, CS and OC only run in a linear approach through the attribute overlap of the two terms. Besides, Srinivas et al. [24] proposed to use Dice Coefficient and Cosine Similarity because both approaches are the best corpus based measurement. After considering the size of the Wikipedia, we decided to use CC, TS, DC, CS and OC in our experiments instead of EMIM.

First and foremost, we used the English Wikipedia (version 2012-01-01), which contains 3,835 million articles in the corpus. For this paper, we only used the first paragraph of each article from the Wikipedia to generate the Term Similarity Matrix because the first paragraph of each article in the Wikipedia pertains the most critical idea of an article and it can stand on it owns as a concise version of this article according to the guideline from Wikipedia. Basically, we generated around 296 million pairs of terms based on the first paragraph of each article from the Wikipedia and Table 1 shows a sample term and values by using DC and CS. In this paper, we will show that different Term Similarity Matrix can impact the retrieval performance. One of the advantages of our approach is that we can generate the Term Similarity Matrix from different types of external resources such as Wikipedia, Dictionary and Thesaurus as long as it is text-based collection.

Table 1.
Sample Term using DC and CS. Note that Words are Stemmed and q refer to Term from Query

Sim (syriah, wj)	DC	CS
Sim(q, wj = assyrian)	0.1797	0.1839
Sim(q, wj = chaldean)	0.1291	0.1470
Sim(q, wj = ephrem)	0.0586	0.1150
Sim(q, wj = nestorian)	0.0833	0.1079
Sim(q, wj = syrian)	0.0852	0.1011
Sim(q, wj = antioch)	0.0971	0.0988
Sim(q, wj = edessa)	0.0722	0.0959
Sim(q, wj = patriarch)	0.0672	0.0853
Sim(q, wj = maronit)	0.0726	0.0803
Sim(q, wj = coptic)	0.0715	0.0715

VI. Experiments

We use CHiC 2012 to test our proposed idea. CHiC 2012 contains fifty queries and one million documents. The uniqueness of this collection is in the average mean of the length of the query, which is 2.84. In this collection, the metadata inside the documents is quite various from large to limited data. Besides, we used external resources such as Wikipedia to generate the Term Similarity Matrix. The proposed model is a generic solution to all application domains. However, CHiC 2012 was chosen as our test collection because the proposed model is more dedicated to the subject of heritage. By using CHiC, the proposed model returns best results and thus, it could be a good benchmark when this generic model is applied to another application domains. In the experiments, we only use the title without any description from the queries. All the experiments were done by using the XIOTA engine [5]. The performance was measured by Mean Average Precision (MAP). The optimal value for Dirichlet prior smoothing for baseline is 100 and 350 for all the Extended Dirichlet. Besides, we applied student's paired t-test (at the $p < 0.06$) to assess the significance of the difference measurement between the several types of statistic approach.

Table 2

Performance with Various Type of Statistic from the First Paragraph of the Articles from Wikipedia (*=Statistical Significance at p<0.06 using the Student's Paired T-Test)

TYPES OF APPROACHES	MEAN AVERAGE PRECISION (MAP)	MAP GAIN OR LOST
BL (baseline): LM with Dirichlet Smooting	0.5273	
LMED-Conf: LM with Extended Dirichlet and CC	0.5196	-1.48%
LMED-T: LM with Extended Dirichlet and TC	0.5395	+2.31%
LMED-D: LM with Extended Dirichlet and DC	**0.5451***	**+3.38%**
LMED-Cos: LM with Extended Dirichlet and CS	0.5435	+3.07%
LMED-O: LM with Extended Dirichlet and OC	0.4929	-6.97%

Table 2 shows clearly that our approach outperforms the baseline result. The most statistical significant improvement is with the LMED-D from 0.5273 to 0.5450 while the most depreciation is with the LMED-O. The reason to these bad results for (LMED-O) is that most of the non-null values of the similarity matrix equal to “1”, which is abnormal because the value of “1” should represent the exact match. Overall, 16 queries show increments, 8 queries show fluctuations and 11 queries remain the same by using LMED-D.

VII. Conclusions and Future Work

We have presented a model to exploit the term similarity of non-matching terms during the retrieval time. Our experiment results indicate that the proposed approach, which is Term Similarity Matrix based on the statistical approach is more efficient and effective than the term intersection approach. For future work, we would like to compute more Term Similarity Matrix from other external resources and not only limited to Wikipedia. If we have more Term Similarity Matrix from different resources, it means we have higher degree of knowledge to build the link between two different terms.

Acknowledgment

This research is fully supported by RAGS from Kementerian Pendidikan Tinggi under the grant number of 2015-0184-109-72.

References

[1] G. Amati and C. J. van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. ACM Transaction on Information Systems, 20(4):357–389, 2002.

[2] M. Bendersky and W. B. Croft. Discovering key concepts in verbose queries. International ACM SIGIR Conference, pp. 491–498, 2008.

[3] A. Berger and J. Lafferty. Information retrieval as statistical translation. International ACM SIGIR Conference, pp. 222–229, 1999.

[4] D. Carmel, N. Zwerdling, and S. Yogev. Entity oriented search and exploration for cultural heritage collections: the EU-Cultura project. International conference companion on World Wide Web, pp. 227–230, 2012.

[5] J.-P. Chevallet. X-iota: An open xml framework for ir experimentation. Lecture Notes in Computer Science, vol. 8, pp. 263–280, 2005.

[6] P. Clough, N. Ford, and M. Stevenson. Personalizing access to cultural heritage collections using pathways. International Workshop on Personalized Access to Cultural Heritage, 2011.

[7] F. Crestani. Exploiting the similarity of non-matching terms at retrieval time. Journal of Information Retrieval, 2:25–45, 2000.

[8] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6):391–407, 1990.

[9] Y. Jing and W. B. Croft. An association thesaurus for information retrieval. In RIAO Conference Proceedings, pages 146–160, 1994.

[10] M. Karimzadehgan and C. Zhai. Estimation of statistical translation models based on mutual information for ad hoc information retrieval. International ACM SIGIR conference, pp. 323–330, 2010.

[11] R. Krovetz. Viewing morphology as an inference process, pp. 191–202, 1993.

[12] V. Lavrenko and W. B. Croft. Relevance based language models. International ACM SIGIR, pp. 120–127, 2001.

[13] C. D. Manning, P. Raghavan, and H. Schutze. Introduction to Information Retrieval. Cambridge University Press, New York, 2008.

[14] B. Markines, C. Cattuto, F. Menczer, D. Benz, A. Hotho, and G. Stumme. Evaluating similarity measures for emergent semantics of social tagging. International Conference on World Wide Web, pp. 641–650, 2009.

[15] H. J. Peat and P. Willett. The limitations of term co-occurrence data for query expansion in document retrieval systems. Journal of the American Society for Information Science, 42:378–383, 1991.

[16] F. Peng, N. Ahmed, X. Li, and Y. Lu. Context sensitive stemming for web search. International ACM SIGIR conference, pp. 639–646, 2007.

[17] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. International ACM SIGIR conference, pp. 275–281, 1998.

[18] M. F. Porter. Readings in information retrieval: An algorithm for suffix stripping, pp. 313–316. Morgan Kaufmann Publishers Inc., 1997.

[19] C. J. V. Rijsbergen. Information Retrieval. Butterworth-Heinemann, Newton, MA, USA, 2nd edition, 1979.

[20] S. E. Robertson. Overview of the okapi projects. Journal of Documentation, 53(1):3–7, 1997.

[21] G. Salton, editor. The SMART Retrieval System Experiments in Automatic Document Processing. Prentice Hall, 1971.

[22] G. Salton. The smart project in automatic document retrieval. International ACM SIGIR conference, pp. 356-358, 1991.

[23] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. In Information Processing And Management, pages 513–523, 1988.

[24] G. Srinivas, N. Tandon, and V. Varma. A weighted tag similarity measure based on a collaborative weight model. International workshop on Search and mining user-generated contents, pp. 79–86, 2010.

[25] J. Xu and W. B. Croft. Query expansion using local and global document analysis. International ACM SIGIR Conference, pp. 4-11, 1996.

[26] L. Zhao and J. Callan. Automatic term mismatch diagnosis for selective query expansion. International ACM SIGIR conference, pp. 515-524, 2012

[27] S. Yogev, H. Roitman, D. Carmel, N. Zwerdling. Towards expressive exploratory search over entity-relationship data. International conference companion on World Wide Web, pp. 83-92, 2012.

[28] C. Zhai, J. Lafferty. A study of smoothing methods for language models applied to information retrieval. ACM Transaction Information System, pp. 179-214, 2004.