Multimedia Information Modeling and Retrieval(MRIM)/Laboratoire d'Informatique de Grenoble (LIG) at CHiC2013

Kian Lam Tan*, Mohannad ALMasri*, Jean-Pierre Chevallet**, Philippe Mulhem***, and Catherine Berrut*

* UJF-Grenoble 1, ** UPMF-Grenoble 2,***CNRS , LIG laboratory, MRIM group
{Kian-Lam.Tan,Mohannad.Almasri,Jean-Pierre.Chevallet,
Philippe.Mulhem,Catherine.Berrut}@imag.fr

Abstract. Numerous cultural heritage materials are accessible through online digital library portals. However, this conversion resulted in the issues of inconsistency and incompleteness. The Cultural Heritage in CLEF 2013 (CHiC) takes the initiative to organize an evaluation campaign which involve several tasks such as 1) multilingual task, 2) polish task and 3) interactive task. We present the results of the MRIM/LIG team for the Ad-Hoc task and for the Semantic Enrichment task. For the Ad-Hoc task, we incorporate Term Links based on Wikipedia into the Language Model. Our approach has the following advantages: 1) it is easy and simple to generate the Term Similarity Matrix based on statistical information 2) a light weight integration in the Language Model. For the semantic query enrichment task, we deal with short queries found in this collection. These short queries can not describe a specific information need. Hence, the goal of this task is to find best ten terms for a query to semantically enrich the topic and guess the user's information need or original query intent. We use the Wikipedia as a semantic resource in order to find these related terms.

Keywords: Information Retrieval, Language Model, Query Enrichment, Query Expansion, Semantic Resource

1 Introduction

Cultural heritage is an expression of the ways of living developed by a community and passed on from generation to generation, including customs, practices, places, objects, artistic expressions and values. Basically, cultural heritage can be distinguished in two types such as artifacts and built environment. Artifacts consist of books, objects, documents and pictures such as Mona Lisa portrait that display at Musee du Louvre, Paris and The Last Supper painting that display at Santa Maria delle Grazie, Milan by Leonardo da Vinci.

Basically, Europeana provides the flexibility for all the people around the world to access the information of cultural heritage such as text, image, audio and video. Therefore, Cultural Heritage in CLEF (CHiC) takes the initiative to

organize the evaluation lab since 2012 to address the key problem from Europeana.

We participated in the English monolingual ad-hoc retrieval task and English monolingual semantic enrichment task.

2 Ad-hoc Retrieval Task

This is a standard ad-hoc retrieval task, which measure the effectiveness of the Information Retrieval System (IRS). The ad-hoc task is the standard setting for IRS which returns a relevance-ranked list of documents based on the query and the collection of the documents.

2.1 Approach

The main idea of this approach is to integrate the term links into the current Dirichlet formula. Firstly, we assume that a term w is $w' \in d$ which can play the role of w where w is $w \in q$ during the matching process. More specifically, we consider that if w does not occur in the initial document d, but it occurs in the *document* d_{ext}, which is the result of the extension of d according to the query and some knowledge [1]. Then, the probability of the term will define according to the extended document d_{ext}.

The knowledge assumes to form a symmetrical similarity function which is $Sim : V \times V \rightarrow [0, 1]$, that denotes the strength of the similarity between two terms from the vocabulary (the larger the value, the higher the strength). We propose that: $\forall w \in V, Sim(w, w') = 1$ if exact matching between w with w', and $\forall w \in V, Sim(w, w') = 0$ if w does not contain any link with w'.

To achieve this, we use 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 $w'' = argmax_{w' \in d, w' \neq w} Sim(w, w')$ as the term from the document will serve as the basis count of the pseudo occurrences of w in d as $c(w''; d).Sim(w'', w)$. 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 contains any link, then it's occurrences is counted in the extended document.

Eventually, using 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:

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

[1] The knowledge refers to the term links

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

$$w'' = argmax_{w' \in d, w' \neq w} Sim(w, w') \tag{1}$$

Using the fact above, the expression of $|d_{ext}|$ can be easily simplified into:

$$|d_{ext}| = |d| + \sum_{w'' \in d \setminus q; Sim(w,w'') \neq 0} c(w''; d).Sim(w'', w) \tag{2}$$

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 d_{ext} according to a query q, noted that $p_\mu(w|d_{ext})$ is defined as:

1. if $w \in d \cap q$:
$$P_\mu(w|d_{ext}) = \frac{c(w; d) + \mu P(w'|C)}{|d_{ext}| + \mu} \tag{3}$$
2. if $\exists w'' \in d \setminus q; Sim(w, w'') \neq 0$:
$$P_\mu(w|d_{ext}) = \frac{c(w"; d).Sim(w, w") + \mu P(w"|C)}{|d_{ext}| + \mu} \tag{4}$$
with $w'' = argmax_{w' \in d, w' \neq w} Sim(w, w')$.
3. if $\nexists w'' \in d \setminus q; Sim(w, w'') \neq 0$
$$P_\mu(w|d_{ext}) = \frac{c(w; d) + \mu P(w|C)}{|d_{ext}| + \mu} \tag{5}$$
with $w'' = argmax_{w' \in d, w' \neq w} Sim(w, w')$.

In the 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 $p_\mu(w|d) = p_\mu(w|d_{ext})$.

2.2 Term Links

Basically, we make the assumption that two terms are considered link to each other if both terms co-occur in the same context. So, the term links contains the link between the term w and w'. In this experiment, we only used Cosine Similarity (CS) to generate the term links. The DC between term w and w' are calculated as follows:

The CS between term w and w' is represented using a dot product and magnitude as follows:

$$Sim_{cosine}(w, w') = \sqrt{\frac{n(w \cap w')}{n(w).n(w')}} \tag{6}$$

Table 1. MAP for the ad-hoc experiments.

Types of Approaches	MAP
LMED-Cos-TL1	0.06340
LMED-Cos-TL2	0.06430

2.3 Experiment, Result and Discussion

All the experiments are done by using the XIOTA engine [3]. The performance is 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 only use the title without any description form the queries and index the title, subject, and description from the documents (CHiC collection). As for pre-processing, we remove all the stop words which contains 571 words and non-character, and apply the Porter Stemming method. On the other hand, we convert all the upper case to lower case. In addition, we use the English Wikipedia (version 2012-01-01) which contains 3.835 million articles to generate the two types of Term Links (we called it as "TermLinks1" and "TermLinks2") based on Cosine Similarity (6). We do not apply Porter Stemming method on "TermLinks1" while we apply Porter Stemming method on "TermLinks2".

The approaches used for the experiments in the following section are:

- LMED-Cos-TL1: LM with Extended Dirichlet, CS, and TermLinks1
- LMED-Cos-TL2: LM with Extended Dirichlet, CS, and TermLinks2

We only submitted two results (since we participated in the English monolingual ad-hoc task) based on our propose approach. Table 1 shows the MAP for the the ad-hoc experiments. Basically, we achieved the highest MAP if we compare to others in the English monolingual ad-hoc retrieval task. Besides, both of our results (LMED-Cos-TL1 and LMED-Cos-TL2) outperforms the rest of the participants in the English multilingual ad-hoc retrieval task except the team from Chemnitx University of Technology, Germany.

3 Semantic Query Enrichment

In this part, we address short queries in ChiC collection which have no sufficient information to express its semantic. For example, assume the query "last supper". A retrieval model will retrieve documents which contain these two words or one of them without any attention to the meaning of this query in the Christian religion. Whereas, if we know this information, some related terms to this meaning like "Jesus", "crucifixion", "twelve apostles", and "Judas" could be found. Then, we can enrich the original query using these related terms. Therefore, the ability of an IRS to retrieve the relevant document to this query can be enhanced. Semantic query enrichment is to find and add these terms which are semantically related to a query. These added terms provide a semantic context

for a query. This context is used by IRS to enhance its relevance estimation in its retrieval task.

Pseudo-Relevance Feedback is one of the most popular methods for finding these enrichment terms using the top k retrieved document to the original query. Whereas, if top retrieved documents for a given query contains a few number of relevant document. In this case, selected terms using Pseudo-Relevance Feedback will not be strongly related to the original query and will introduce noise into the enriched query. As a result, the relevance estimation for the enriched query would be less or equal than the original query [4, 2, 1].

We present another method in order to select related terms for a given query using an external knowledge. Many resources are available in order to achieve this task: ontologies, encyclopedias, lexical resources. We use, in our task, Wikipedia as an external knowledge in order to achieve semantic query enrichment. Given a query q, in our case, this query talks about one well known thing: person, place, event, etc.. Wikipedia is a freely available large knowledge which contains a huge number of articles and links between them. First, we present the structure of Wikipedia. Then, we present our semantic query enrichment approach which is based on this structure.

3.1 Wikipedia Structure

Wikipedia is a knowledge base which can be represented as a directed weighted graph of articles. The basic entry in Wikipedia is an entity page, which is an article that contains information focusing on one single entity. Furthermore, each article is linked to other articles by a number of weighted links. This weights represent how much the two entities are semantically related. An article point to a collection of articles and is pointed by a collection of other articles Figure 1.

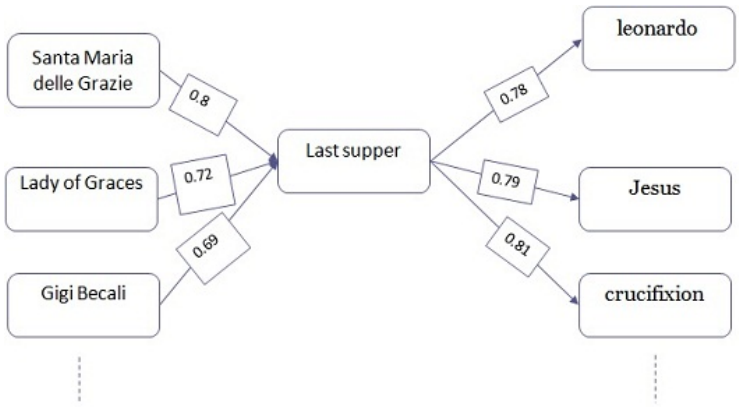

Fig. 1. Example of Wikipedia article and its relation with other articles

3.2 Enrichment Steps

As we mentioned before, our semantic query enrichment use Wikipedia as a knowledge base. We see from the previous section, Wikipedia is organized as directed weighted graph of articles. Each article is identified by its title, links in, and links out. Using Wikipedia, each text can be mapped into a collection of articles. Relaying on what mentioned about Wikipedia, we present our semantic query enrichment steps:

- Given a query, first, finding all articles which correspond this query in Wikipedia, we call them: identified articles.
- Using the identified articles we have different variants to enrich the original query q:
  - o Links in: candidate articles to enrich the original query, in this first case, all articles which point out to at least one article of the identified articles.
  - o Links out: candidate articles to enrich the original query, in this second case, all articles which are pointed out by at least one article of the identified articles.
  - o Mixed: candidate articles to enrich the original query, in this last case, contain the union between articles form first and second case.
- Sort candidate articles depending on its relatedness to the identified concepts.
- Take best k articles titles from candidate articles and add them to the original query.
- For weighting these articles, we multiplied the relatedness values using different values between [0, 1] like the following (0, 0.1, 0.2, 0.3, ,1). The value which provided the best precision enhancement was 0.3.

Using these steps, we obtain best k related titles to a given query with their wights. These titles are added to this query to obtain a long query. We claim that this long query has sufficient information to express the information need. Therefore, it is proposed to help IRS to enhance its relevance estimation or in other words its precision.

3.3 Experiment and Result

Experiments are done using WikipediaMiner[2] which is an API for searching and accessing Wikipedia content. We mean by content articles and their links. WikipediaMiner is a toolkit for tapping the rich semantics encoded within Wikipedia. It helps to integrate Wikipedia's knowledge into applications, by:1) providing simplified, object-oriented access to Wikipedia's structure and content.2) Measuring how terms and concepts in Wikipedia are semantically related to each other.

[2] http://wikipedia-miner.cms.waikato.ac.nz/index.html

We validate our approach over CHIC2013 English collection. For the query enrichment task we have 25 queries. These queries contain well known entities like persons, events, etc. The task requires systems to present a ranked list of at most 10 related terms for a query to semantically enrich the topic and/or guess the user's information need or original query intent. Related terms in our case are extracted using WikipediaMiner.

The evaluation metric for the semantic enrichment task is precision (precision@1, @3, @10) Table 2. Precision at a given index k measure if the first k enrichment terms to a given query are related to this query or not. In this table, we have two runs, in first run we use in enrichment a mix between links in and links out. We select the 5 top articles titles form link in and the 5 top articles titles from link out. In the second run, we use best 10 articles titles from links out (best means most semantically related depending on Wikipedia relatedness values between Wikipedia articles). Basically, our second result (MRIM_SE13_EN_WM_1) outperforms the other participants for monolingual English enrichment by means $P@1$ and $P@3$. Whereas, it is slightly less of them by means of $P@10$.

Table 2. Semantic enrichment task results(precision@1, @3, @10)

Run Name	$P@1$	$P@3$	$P@10$
MRIM_SE13_EN_WM	0.2800	0.1333	0.1448
MRIM_SE13_EN_WM_1	0.2800	0.1467	0.1598

4 Conclusion

For the ad-hoc retrieval task, our results indicated that both results (LMED-Cos-TL1 and LMED-Cos-TL2) achieved almost the same MAP. Based on this scenario, we can conclude that there is not much different to apply Porter Stemming method on the Term Links since the gap between these two results is very small. Whereas, in the semantic enrichment task, our results show that using links out is better of using the mix between links in and out.

References

1. Eneko Agirre, Paul D. Clough, Samuel Fernando, Mark Hall, Arantxa Otegi, and Mark Stevenson. The sheffield and basque country universities entry to chic: Using random walks and similarity to access cultural heritage. In *CLEF (Online Working Notes/Labs/Workshop)*, 2012.
2. Mitra Akasereh, Nada Naji, and Jacques Savoy. Unine at clef 2012. In *CLEF (Online Working Notes/Labs/Workshop)*, 2012.

3. Jean-Pierre Chevallet. X-iota: An open xml framework for ir experimentation. In SungHyon Myaeng, Ming Zhou, Kam-Fai Wong, and Hong-Jiang Zhang, editors, *Information Retrieval Technology*, volume 3411 of *Lecture Notes in Computer Science*, pages 263–280. Springer Berlin Heidelberg, 2005.
4. Jinxi Xu and W. Bruce Croft. Query expansion using local and global document analysis. In *In Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 4–11, 1996.