Multi-facet Document Representation and Retrieval
Karam Abdulahhad, Jean-Pierre Chevallet, Catherine Berrut

To cite this version:
Karam Abdulahhad, Jean-Pierre Chevallet, Catherine Berrut. Multi-facet Document Representation and Retrieval. CLEF 2011 - Conference on Multilingual and Multimodal Information Access Evaluation, Sep 2011, Amsterdam, Netherlands. 9p. hal-00764761

HAL Id: hal-00764761
https://hal.archives-ouvertes.fr/hal-00764761
Submitted on 13 Dec 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Multi-facet Document Representation and Retrieval

Karam Abdulahhad*, Jean-Pierre Chevallet**, and Catherine Berrut*

* UJF-Grenoble 1, ** UPMF-Grenoble 2, LIG laboratory, MRIM group
karam.abdulahhad,jean-pierre.chevallet,catherine.berrut@imag.fr

Abstract. This paper presents our participation in ImageCLEF2011, in the two tasks: ad-hoc image-based retrieval and case-based retrieval, of the medical retrieval track.

We participated through a simple IR model based on three hypotheses: 1) the amount of overlap between a document and a query, 2) the descriptive power of an indexing element, and 3) the discriminative power of an indexing element. We used three types of indexing elements: ngrams, keywords, and concepts, for building and checking the effectiveness of multi-facet document representation.

Although of the simplicity of our model in both documents’ representation and retrieval, we could obtain good results. The eighth out of 64 runs in the ad-hoc image-based retrieval task, and the fifth out of 35 runs in the case-based retrieval task.

1 Introduction

Each information retrieval (IR) model has its advantages and drawbacks. In other words, an IR model may perform well in some cases but badly in others. In general, there is no IR model performs well in all cases [9].

Our model is not an exception. Therefore, a general mechanism to evaluate the performance of an IR model and compare it with the performance of others is needed. In this context, there are many campaigns of evaluation in IR field, e.g. CLEF1.

First, our model is a text-based retrieval and very simple model. It uses multiple types of indexing elements, e.g. ngrams, keywords, or concepts. The goals of this research are: 1) studying the performance of the model itself, using one of the indexing element types, 2) studying the effects of using multiple indexing element types at the same time on the performance.

Second, CLEF (Cross-Language Evaluation Forum) is a yearly evaluation campaign in Multilanguage information retrieval field since 2000. ImageCLEF2 is a part of CLEF. It concerns searching medical images through documents that contain text and images [14].

1 http://www.clef-campaign.org/
2 http://www.imageclef.org/
This year, ImageCLEF2011[^3] contains four main tracks: 1) medical retrieval, 2) photo annotation, 3) plant identification, and 4) Wikipedia retrieval. Medical retrieval track contains three tasks: 1) modality classification, 2) ad-hoc image-based retrieval which is an image retrieval task using textual, image or mixed queries, and 3) case-based retrieval: in this task the documents are journal articles extracted from PubMed[^4] and the queries are case descriptions.

Third, we participated in the last two tasks: ad-hoc image-based retrieval and case-based retrieval.

The test collection of this year 2011 contains, according to each task: 1) ad-hoc image-based retrieval: about 230,000 images with their text caption written in English and 30 queries written in three languages: English, French, and German. 2) case-based retrieval: about 55,000 articles written in English and 10 queries written in English.

This paper is structured as follows: Section 2 describes in details our model and the different types of indexing elements that are used. Section 3 presents some technical aspects of applying this model for ImageCLEF2011 test collection. Section 4 discusses the obtained results. We conclude in section 5.

2 The Proposed Model

2.1 Three Types of Indexing Elements

Any IR model should contain two main components: indexing function and matching function. The goal of the indexing function is to convert documents and queries from their original form to another easy to use form.

\[
\text{Index} : D \cup Q \rightarrow E^* \quad (1)
\]

Where
- \(D\) set of documents
- \(Q\) set of queries
- \(E\) set of indexing elements
- \(E^*\) the set of all subsets of \(E\)

Concerning indexing elements, three different types are used: ngrams \((NG)\), keywords \((K)\), and concepts\(^5\) \((C)\). Therefore, three indexing functions are exist (one for each type).

\[
\text{Index}_{NG} : D \cup Q \rightarrow E^*_{NG} \quad (2)
\]

\[
\text{Index}_K : D \cup Q \rightarrow E^*_K \quad (3)
\]

[^5]: "Concepts" can be defined as "Human understandable unique abstract notions independent from any direct material support, independent from any language or information representation, and used to organize perception and knowledge" [8]. In IR domain, to achieve the conceptual indexing, each concept is associated to a set of terms that describe it [3] [7].
\[ \text{Index}_C : D \cup Q \rightarrow E_C^* \] (4)

Where
\( E_{NG} \) set of ngrams
\( E_K \) set of keywords
\( E_C \) set of concepts

We believe that no single type of indexing elements could completely represent the content of documents and queries, because: 1) there is no perfect indexing function [3] [2] [11]. It is always an approximate function, 2) concerning concepts, the most of resources that contain concepts, e.g. UMLS\(^6\), are incomplete [5] [6] [1], 3) each type covers an aspect of documents and queries [10]. Ngrams cover the statistical aspect, keywords cover the lexical aspect, and concepts cover the conceptual aspect.

### 2.2 Matching Function

Our model, as almost all models, depends on some hypotheses. Actually, it depends on the following three hypotheses:

1. The more shared elements a document and a query have, the semantically closer they are.
2. The descriptive power of an element (local weight): the more frequently an element occurs in a document, the better it describes the document [13] [3].
3. The discriminative power of an element (global weight): the less number of documents an element appears in, the more important it is [13] [3].

By taking these hypotheses into account, our model could be formulated. For any type of indexing elements the Relevance Status Value (RSV) between a document \( d \) and a query \( q \) is:

\[
\text{RSV} (d, q) = \|d \cap q\| \times \left( \sum_{e \in q} \frac{N_e}{N} \times \frac{f_{d,e}}{\|d\|} \right)
\] (5)

Where
\( d = \{e|e \in \text{Index} (d)\} \) a document
\( q = \{e|e \in \text{Index} (q)\} \) a query
\( \|d \cap q\| = ||\{e|e \in \text{Index} (d) \cap \text{Index} (q)\}|| \) the number of shared elements between a document \( d \) and a query \( q \)
\( N \) the number of documents in the corpus
\( N_e = ||\{d|e \in \text{Index} (d)\}|| \) the number of documents that contain the element \( e \)
\( f_{d,e} \) the number of occurrences of an element \( e \) in a document \( d \)
\( \|d\| \) the number of elements in a document \( d \)

\(^6\) Unified Medical Language System. It is a meta-thesaurus in medical domain.
2.3 Matching Function According to each Type of Indexing Elements

Now after presenting our model in general, it should be instantiated according to each type of indexing elements.

**Ngrams**

\[
RSV_{NG}(d, q) = \|d \cap q\|_{NG} \times \left( \frac{\sum_{ng \in q} \frac{N}{N_{ng}} \times f_{d,ng}}{\|d\|_{NG}} \right)
\]  

(6)

Where

- \(d = \{ng|ng \in Index_{NG}(d)\}\) a document
- \(q = \{ng|ng \in Index_{NG}(q)\}\) a query
- \(\|d \cap q\|_{NG} = \|\{ng|ng \in Index_{NG}(d) \cap Index_{NG}(q)\}\|\) the number of shared ngrams between a document \(d\) and a query \(q\)
- \(N\) the number of documents in the corpus
- \(N_{ng} = \|\{d|ng \in Index_{NG}(d)\}\|\) the number of documents that contain the ngram \(ng\)
- \(f_{d,ng}\) the number of occurrences of a ngram \(ng\) in a document \(d\)
- \(\|d\|_{NG}\) the number of ngrams in a document \(d\)

**Keywords**

\[
RSV_{K}(d, q) = \|d \cap q\|_{K} \times \left( \frac{\sum_{k \in q} \frac{N}{N_{k}} \times f_{d,k}}{\|d\|_{K} \times \|k\|} \right)
\]  

(7)

Where

- \(\|k\|\) the number of characters in a keywords \(k\)

**Concepts**

\[
RSV_{C}(d, q) = \|d \cap q\|_{C} \times \left( \frac{\sum_{c \in q} \frac{N}{N_{c}} \times f_{d,c}}{\|d\|_{C}} \right)
\]  

(8)

2.4 The Three Types in one Model

As we said earlier, no single type of indexing elements could cover all aspects of documents and queries. Therefore, merging all types (aspects) in one model could enhance the performance of our model [9] [4] [16]. One of the merging formulas is:

\[
RSV_{all}(d, q) = RSV_{NG}(d, q) + RSV_{K}(d, q) + RSV_{C}(d, q)
\]  

(9)

To increase the chance of retrieving more documents, another component could be added to the Formula (9). The component represents an expansion
of the query q. It is $\|d \cap q\|_{\text{expan}}$: the number of shared keywords between a document d and a query q after replacing each query’s concept c that does not occur in d by the set of keywords that represent $c^7$.

$$RSV_{all}(d, q) = RSV_{NG}(d, q) + RSV_K(d, q) + RSV_C(d, q) + \|d \cap q\|_{\text{expan}}$$ (10)

3 Model Validation

3.1 Ad-hoc Image-Based Retrieval

In this task, image captions are used as documents and the English part of queries is just taken into account.

Text indexing: we extracted three types of indexing elements:

1. 5gram$^8$: before extracting 5grams from documents and queries, we deleted all non-ASCII characters. Then we used five-characters-wide window for extracting 5grams with shifting the window one character each time.

2. Keywords: before extracting keywords from documents and queries, we deleted all non-ASCII characters. Then we eliminated the stop words and stemmed the remaining keywords using Porter algorithm to get finally the list of keywords that index documents and queries.

3. Concepts: before mapping the text of documents and queries to concepts, we deleted all non-ASCII characters. Then we mapped the text to UMLS’s concepts using MetaMap [2].

Model variants: actually we experimented five variants of our model in this task, which are:

$$RSV_{all}(d, q) = RSV_{NG}(d, q) + RSV_K(d, q) + RSV_C(d, q)$$ (11)

$$RSV_{all}(d, q) = (\|d \cap q\|_{5G,K,C}) \times (\text{sum}_{5G,K,C})$$ (12)

$$RSV_{all}(d, q) = RSV_{NG}(d, q) + RSV_K(d, q) + RSV_C(d, q) + \|d \cap q\|_{\text{expan}}$$ (13)

Where

$$\|d \cap q\|_{5G,K,C} = \|d \cap q\|_{5G} + \|d \cap q\|_{K} + \|d \cap q\|_{C}$$

$$\text{sum}_{5G,K,C} = \text{sum}_{5G} + \text{sum}_{K} + \text{sum}_{C}$$

$$\text{sum}_{5G} = \left(\sum_{5g \in q} \frac{N_{5g}}{N_c} \times \frac{f_{d,5g}}{\|d\|_{5G}}\right)$$

$$\text{sum}_{K} = \left(\sum_{k \in q} \frac{\|k\|}{\|d\|_{K}}\right)$$

$$\text{sum}_{C} = \left(\sum_{c \in q} \frac{\|c\|}{\|d\|_{C}}\right)$$

$$RSV_{all}(d, q) = (\|d \cap q\|_{5G,K,C}) \times (\text{sum}_{5G,K,C} + \|d \cap q\|_{\text{expan}})$$ (14)

$$RSV_{all}(d, q) = ((\|d \cap q\|_{5G,K,C}) \times (\text{sum}_{5G,K,C})) + \|d \cap q\|_{\text{expan}}$$ (15)

$^7$ We supposed that each concept is represented by a set of keywords in its resource (we used UMLS as resource).

$^8$ 5gram is a ngram consists of five characters. We picked out 5grams because they gave the best results using ImageCLEF2010 comparing to the other ngrams.
3.2 Case-Based Retrieval

In this task, articles are used as documents. We indexed documents and queries in the same way as in the previous task. However, we extracted only two types of indexing elements (4grams, keywords) because of technical reasons.

**Model variants**: actually we experimented two variants of our model in this task, which are:

\[
RSV_{all} = RSV_{4G} + RSV_K
\]  

\[
RSV_{all}(d,q) = (\|d \cap q\|_{4G,K}) \times (sum_{4G,K})
\]

Where

\[
\|d \cap q\|_{4G,K} = \|d \cap q\|_{4G} + \|d \cap q\|_{K}
\]

\[
sum_{4G,K} = sum_{4G} + sum_{K}
\]

\[
sum_{4G} = \left( \sum_{4g \in q} \frac{N_{4g}}{N_4} \times \frac{f_{4g,4}}{\|d\|_{4G}} \right)
\]

\[
sum_{K} = \left( \sum_{k \in q} \frac{N_{k}}{N_K} \times \frac{f_{k,K}}{\|d\|_{K}} \right)
\]

4 Results and Discussion

Before we start representing and discussing the obtained results, we will present the names that we will use in our discussion for referring to each variant with its corresponding formula and run’s name that used in the official campaign\(^9\) (see Table 1).

<table>
<thead>
<tr>
<th>Variant’s name</th>
<th>Corresponding formula</th>
<th>Corresponding run’s name</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICK5G_1</td>
<td>Formula 11</td>
<td>IVSCT5G</td>
</tr>
<tr>
<td>ICK5G_2</td>
<td>Formula 12</td>
<td>IVPCT5G</td>
</tr>
<tr>
<td>ICK5G_Q1</td>
<td>Formula 13</td>
<td>IVSCT5GK</td>
</tr>
<tr>
<td>ICK5G_Q2</td>
<td>Formula 14</td>
<td>IVPCT5GKm</td>
</tr>
<tr>
<td>ICK5G_Q3</td>
<td>Formula 15</td>
<td>IVPCT5GKout</td>
</tr>
<tr>
<td>CK5G_1</td>
<td>Formula 16</td>
<td>MRIM,KI_A,VM_Sop,T4G</td>
</tr>
<tr>
<td>CK5G_2</td>
<td>Formula 17</td>
<td>MRIM,KI_A,VM_Pos,T4G</td>
</tr>
</tbody>
</table>

4.1 Ad-hoc Image-Based Retrieval

The following table (see Table 2) contains the obtained results. The first row (Best) is the result of the first ranked run in the ad-hoc image-based retrieval task.

From one side, although of using a very simple structure (set of elements) to represent the content of documents and queries, and using a simple formula

\(^9\) To see all results: http://www.imageclef.org/2011/medical
Table 2. The results of ad-hoc image-based retrieval task

<table>
<thead>
<tr>
<th></th>
<th>MAP</th>
<th>P@10</th>
<th>P@20</th>
<th># rel</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best</td>
<td>0.2172</td>
<td>0.3467</td>
<td>0.3017</td>
<td>1471</td>
<td>1</td>
</tr>
<tr>
<td>ICK5G1</td>
<td>0.2008</td>
<td>0.3033</td>
<td>0.3050</td>
<td>1544</td>
<td>8</td>
</tr>
<tr>
<td>ICK5GQ1</td>
<td>0.2008</td>
<td>0.3033</td>
<td>0.3050</td>
<td>1543</td>
<td>9</td>
</tr>
<tr>
<td>ICK5GQ2</td>
<td>0.1975</td>
<td>0.2967</td>
<td>0.2833</td>
<td>1517</td>
<td>10</td>
</tr>
<tr>
<td>ICK5G2</td>
<td>0.1974</td>
<td>0.2967</td>
<td>0.2833</td>
<td>1520</td>
<td>11</td>
</tr>
<tr>
<td>ICK5GQ3</td>
<td>0.1973</td>
<td>0.2967</td>
<td>0.2833</td>
<td>1519</td>
<td>12</td>
</tr>
</tbody>
</table>

(see Formula 5) to compute the matching value between a document and a query, we obtained good results. The run $ICK5G_1$ is ranked eighth out of 64 runs. That’s because, we explicitly represented multi-facet (multi aspects) of documents and queries. In other words, three types of elements: 5Grams, Keywords, and Concepts are used and involved in the matching process.

From another side, using different fusion formulas to merge the results of using different types of indexing elements does not change a lot of things. Compare the results of the two runs $ICK5G_1$ and $ICK5G_2$ (see Table 2).

In addition, the query expansion and the different formulas to integrate it into the model did not add anything. Compare the results of the three runs: $ICK5G_{Q1}$, $ICK5G_{Q2}$, and $ICK5G_{Q3}$ with the result of the run $ICK5G_1$ (see Table 2). That’s because, the added value of query expansion is already compensated by using keywords and 5Grams.

Another noted result is that our model could retrieve the most relevant documents comparing to the other runs.

4.2 Case-Based Retrieval

The following table (see Table 3) contains the obtained results. The first row (Best) is the result of the first ranked run in the case-based retrieval task.

Table 3. The results of case-based retrieval task

<table>
<thead>
<tr>
<th></th>
<th>MAP</th>
<th>P@10</th>
<th>P@20</th>
<th># rel</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best</td>
<td>0.1297</td>
<td>0.1889</td>
<td>0.1500</td>
<td>144</td>
<td>1</td>
</tr>
<tr>
<td>IC_K4G1</td>
<td>0.1114</td>
<td>0.1444</td>
<td>0.1444</td>
<td>149</td>
<td>5</td>
</tr>
<tr>
<td>IC_K4G2</td>
<td>0.0911</td>
<td>0.1111</td>
<td>0.1278</td>
<td>146</td>
<td>10</td>
</tr>
</tbody>
</table>

Here also, we obtained good results. The run $C_K4G_1$ is ranked fifth out of 35 runs, knowing that, we used a very simple structure (set of elements), a

\[http://www.imageclef.org/2011/medical\]
simple matching formula (see Formula 5), and also two simple types of indexing elements: Keywords and 4Grams. The two-facet (Keywords and 4Grams) representation of documents and queries was useful.

However, the formulas of merging the results of using different types of indexing elements were more sensitive comparing to the formulas in ad-hoc image-based retrieval task. Compare the results of the two runs: C_K4G_1 and C_K4G_2 (see Table 3). That’s maybe because, we used less number of elements’ types, two types (Keywords and 4Grams) in case-based retrieval comparing to three types (Concepts, Keywords, and 5Grams) in ad-hoc image-based retrieval.

5 Conclusion

We presented in this paper our approach to index and retrieve documents. We used three types of indexing elements (ngrams, keywords, concepts) for building a multi-facet document representation, and then we used a simple formula based on three hypotheses (the amount of overlap between a document and a query, the descriptive power of an indexing element, and the discriminative power of an indexing element) for retrieving documents, considering all facets (elements’ types) of documents.

We obtained good results. The eighth out of 64 runs in the ad-hoc image-based retrieval task, and the fifth out of 35 runs in the case-based retrieval task, knowing that, we used a very simple structure for representing documents and queries and also a very simple ranking formula.

References


