

# IPAL at ImageClef 2007

## Mixing Features, Models and Knowledge.

Sheng Gao, Jean-Pierre Chevallet, Thi Hoang Diem Le, Trong Ton Pham, Joo Hwee Lim  
IPAL French-Singaporean Joint Lab  
Institute for Infocomm Research (I2R)  
Centre National de la Recherche Scientifique (CNRS)  
21 Heng Mui Keng Terrace - Singapore 119613  
`gaosheng,viscjp,stuthdl,ttpham,jooHwee@i2r.a-star.edu.sg`

### Abstract

This paper presents IPAL ad-hoc photographic retrieval and medical image retrieval results in the ImageClef 2007 campaign. For the photo task, IPAL group is ranked at the 3rd place among 20 participants. The MAP of our best run is 0.2833, which is ranked at the 6th place among the 476 runs. The IPAL system is based on the mixed modality search, i.e. textual and visual modalities. Compare with our results in 2006, our results are significantly enhanced by extracting multiple low-level visual content descriptors and fusing multiple CBIR. Several text based image search (TBIR) engines are also developed such as the language model (LM) approach, the latent semantic indexing (LSI) approach. We also have used external knowledge like Wordnet, and Wikipedia for document expansion. Then the cross-modality pseudo-relevance feedback is applied to boost each individual modality. Linear fusion is used to combine different ranking lists. Combining the CBIR and TBIR outperforms the individual modality search. On medical side, our run ranks 19 among 111. We continue to use a conceptual indexing with vector space weighting, but we add this year a Bayesian network on concepts extracted from UMLS meta-thesaurus.

### Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval—*Query Formulation*

### General Terms

Algorithms, Measurement, Performance, Experimentation

### Keywords

Language model, latent semantic indexing, information retrieval, multimodality fusion, content-based image retrieval, text based image retrieval

## 1 Introduction

This year IPAL group continues to participate both in the ad-hoc photographic retrieval task and the medical task. The photographic retrieval task is different from the task of last year in that this year 1) the notes in the image document are not used while last year they can be used, 2) only text keywords in the title field are used for the query while last year all keywords in all fields

could be used, 3) query image samples are excluded from the image database, and 4) 244 new images are added. Thus there is much less text information available, which obviously decreases discrimination, especially for the query. That is the reason why we have tried this year to expend documents using external information source (Wikipedia).



Figure 1: Images with very similar text content but very different visual content.

After removing stop words, the average query length is about 3.3 text keywords with the maximal 6 keywords and the minimal 2 keywords. The less text keywords also make it impossible to discriminate some image documents which have same text keywords; however, their real visual contents are much different. Some image documents in the database have very similar text decryptions. Thus, only text information cannot distinguish the two images. However, when we measure their visual content, it is easy to discriminate one from another. For example in the figure 1, both images have similar textual content: "Accommodation Huanchaco - Exterior View, Trujillo, Peru" except one word "interior/exterior". The left image ("exterior", image id=01/1311), is relevant with the query 1, i.e. "accommodation with swimming pool", while the image at right ("interior", image id=01/1310) is irrelevant. With some external knowledge one could deduce that an "interior view" has less chance to show a swimming pool. Unfortunately, this remains difficult to do in practice. Without being able to use in an efficient way, extra knowledge for the matching, this example clearly shows that the visual content plays the same important role and can complement text matching. Therefore, this year we pay much attention to improve our Content based Image Retrieval (CBIR) system. We also try to use external knowledge from Wikipedia.

## 2 Photographic Image Indexing

We build various CBIR and Text Based Information Retrieval (TBIR) systems using different indexing methods and similarity functions. Totally we submit 27 runs. In the following, the details are given. To enrich the visual content representation, 4 types of low-level visual feature are extracted from the local regions or global images. They are detailed in the following:

**COR:** Auto color correlogram with the depth 2 in the HSV space. It is extracted from the whole image and is represented by one 324-dimensional vector.

**HSV:** Histogram in HSV and gray-level space with 162-dimension for HSV plus 4-dimension for gray-level. An image is represented by a 166-dimensional vector.

**EDGE:** We used canny operator to detect the contours object. A vector of 80-dimensional is used to capture magnitudes and gradient of the contours.

**GABOR:** texture feature using Gabor filter in the uniformly segmented 5x5 grids at the 2-scale and 12-orientations. Thus, the mean and variance are calculated at each grid to generate 48-dimensional feature vector.

**SIFT:** the SIFT feature is extracted using David Lowe' tool and the region around the keypoint is described by a 128-dimensional appearance feature [8].

As the above low-level features are obtained, we now begin to discuss how to index the image and calculate the similarity score. For the first two visual features extracted from the image, two methods are developed to index the image.

**tf-idf:** the tf-idf method is a popular way to describe the text document [2]. Here we use it to process the histogram. We treat each bin in the histogram as a distinctive word. Then we can calculate the occurrence number of the word occurred in the image, which is the frequency in the bin, and the number of the images contained the word in the database. Now the histogram feature is further processed using the tf-idf method to index the image. The ranking function is the cosine function.

**SVD:** use Single Value Decomposition (SVD) to remove the correlation among the feature components. Assuming the full rank is  $N$ , we empirically select the top  $N*0.8$  eigenvectors and index the image in the eigen-space. The cosine function is used to calculate similarity score.

**Bag-of-visual-words:** images are divided into 16 equal patches. Then, image features are extracted for each patch. Patches are clustered into 1000 clusters using their feature vectors. Finally, we quantize each image by a discrete vector. Image retrieval will be performed by normalizing and matching using their vector signatures.

**ISM:** the ranking function is learned from the training sample using the supervised learning algorithms. Here it is the Integrated Statistic Model (ISM) based binary classifier [4]. The training samples include 3 query images as the positive class and another 3 negative samples randomly selected from the image database. Then the likelihood ratio between the positive model and the negative model is calculated for each image documents in the database and all images are ranked from the highest value to the lowest one.

Therefore, 3 individual runs are generated for each type of visual features. For the other two grid/patch based visual features, i.e. Gabor and SIFT, we use the following approaches to calculate the ranking score:

**HME:** Use the supervised learning approach discussed above to train the hidden maximum entropy (HME) based model using the Gabor or SIFT feature and then use likelihood ratio for ranking images.

**WORD\_SVD:** the method is to index images in word space rather than the image signal level (Gabor or SIFT space). To do so, we first find 200 most frequent keywords by analyzing the attached text documents in the image database. For each keyword, a HME classifier is trained using the corresponding training set. The training set for a word is constructed as follows: if the attached text for an image contains the word, then it is labelled as a positive sample. Otherwise, it is labelled as the negative. After 200 HME classifiers are trained, we will get all likelihood ratio scores for an image to index the image at the 200-dimensional keyword space. To further reduce the correlation among the feature components, SVD is used as discussed above. The similarity function is still the cosine distance.

**WORD\_ISM:** we can also train the ISM model for each query topic using the 200-dimensional feature in the keyword space. Then the likelihood ratio is used as the ranking score.

So there are also 3 runs for the Gabor or SIFT feature to be generated. Totally we have 12 visual runs. Table 2 summarizes the individual visual runs. Row are feature type, column are indexing method. A "+" means the indexing method is available for the feature

## 2.1 Pseudo-relevance Feedback

Learned from the ImageClef 2006 [9], the cross-modality pseudo-relevance feedback (PRF) can improve the system performance, i.e. the TBIR can be boosted by the top- $N$  (here 10 documents are selected) documents from the CBIR as the feedback and vice versa. This year PRF is also adopted.

Table 1: Summary of individual visual runs

|       | tf-idf | SVD | ISM | HME | WORD_SVD | WORD_ISM |
|-------|--------|-----|-----|-----|----------|----------|
| COR   | +      | +   | +   |     |          |          |
| HSV   | +      | +   | +   |     |          |          |
| GABOR |        |     |     | +   | +        | +        |
| SIFT  |        |     |     | +   | +        | +        |

### 3 Text Indexing

We think the use of knowledge to assist the retrieval task is important, because, retrieval is clearly a human task linked to understanding. Knowledge can be defined as: information used to complete a task. In our case it is every piece of information that can be used to help the derivation of the document to the query. For example, the language used, the part of speech of words, and the grammar, are simple examples of knowledge that can be useful to use for text retrieval. The document itself can be viewed as a source of knowledge in different steps of the retrieval process:

- At indexing time: in the matching model. For example, the use of inverse document frequency (idf) is a simple way to use statistical global knowledge from a corpus. In a language model [12], the collection is also used to smooth the language model of the document.
- At querying time: pseudo relevance feed back is a technique that uses content of top documents that are supposed to be relevant, to enlarge the query. In a sense, part of information of the corpus is used to help to interpret the query. Co-occurrence thesaurus is obtained from text-mining from the corpus. It can be used also to enlarge the query. The effect is similar to pseudo-feedback.

In this experiment, we have both used knowledge from document corpus and external knowledge (i.e. information that are not part of documents). The external source we use is Wikipedia. This choice is appropriate because Wikipedia includes a large amount of proper nouns, and specially named entities. In IAPR collection, named entities play a major role.

#### 3.1 Using Language Model and Document Knowledge

We have built two individual runs using the language model (LM) based information retrieval [12] and latent semantic indexing (LSI) method [3] [5]. The attached text document for each image is provided in XML format. It is pre-processed as follows: 1) extract pure text, 2) remove the stop words and 3) do the stemming. For the LM-based TBIR, we first build a lexicon dictionary (6,644 words) from all text documents and then train a unigram language model only based on the attached text document for each image in the database. This is done using the CMU-Cambridge Statistical Language Modelling Toolkit [13] with the Witten-Bell smoothing. We have also produced some run using the Lemur Toolkit<sup>1</sup>. With Lemur we use absolute discount smoothing.

Each image is indexed by the word probabilities in the lexicon. Given a query topic and any image in the database, the probability of query words generated by the corresponding image LM can be calculated. The image documents in the database are ranked by the probability from the highest to the lowest. For the LSI-based TBIR, we first use latent semantic indexing [3, 5] to find the 1,635-dimension eigen space. Then the image text document is indexed by a 1,635-dimensional vector. The cosine distance is applied to get the similarity scores between the query words and the image documents.

---

<sup>1</sup><http://www.lemurproject.org>

## 3.2 Using Language Model and External Knowledge (Wikipedia)

There is an increasing production of information in a collaborative form in the "Wiki" format. It starts from Wikipedia<sup>2</sup>, which is a free Encyclopedia, but there exists also now Wiktionary<sup>3</sup>, a free open dictionary, Wikiquote<sup>4</sup> a compendium of quotations from notable people, Wikibooks and Wikisource, free collections of open-content text books, Wikinews, etc. All these information can be valuable source to help document indexing because of their large coverage. The drawback is the lack of explicit structure: these documents are produced for human reading only.

The full content of Wikipedia can be freely downloaded<sup>5</sup> in XML format. XML is a more convenient form to extract some information than HTML because tags refer to content and not only layout. This format is also more adapted for automatic mining than HTML because of special tags (i.e. non XML) that are used to produce the HTML page. They can be used to localize typed content. For example information boxes on the right of a Wikipedia page, comes from different types of structures: **Infobox**, **Geobox**, **Taxobox**. These are structured in explicit fields: **name**, **type**, etc. They are templates which provide structured standardized information, mainly for named entities.

Figure 2: Example of information box

```
{{Infobox Mountain
| Name=Quilotoa [[Image:Flag_of_Ecuador.svg|38px]]
| Photo=QuilotoaPanorama1.jpg
| Caption=Panorama of the lake-filled Quilotoa caldera.
| Elevation=3,914 m
| Location=[[Ecuador]]
| Range=[[Andes]]
| Prominence =
| Coordinates = {{coord dm|00|85|S|78|90|W|type:mountain_region:EC}}
| Topographic map =
| Type=[[Caldera]]
| Age= approx. 40,000 years
| Last eruption= 1280 (?)
| First ascent=
| Easiest route=
}}
```

We have used the dump of 2 April 2007 of Wikipedia in English, which is a 9.8Gb file. This file is organized in 4881983 pages. We have used the **Taxobox** structure to extract a list of animal and plants. This structure is used to abstract common fact in a box about taxonomy of living things (animal and plants). We found 36188 of this box from which we extract 34321 terms, using the **name** and **regnum** field. For CLEF, only the 23399 animal terms are useful for queries 5, 20 and 35. Due to lack of disambiguation, terms like *wall* (a butterfly) and *plane* (a tree) can cause noise.

We have also used the **Geobox** for geographical terms. There is only 1805 box of this type, from which we can extract only 709 terms because we put a strong typographical constraint on terms, for example, we exclude abbreviations. This list contains **river**, **mountain**, **region**, **city**. For CLEF only mountain should be useful for queries 4, 44. The general box type **Infobox** is the most common in this version of Wikipedia, with 320311 objects. The keyword **Infobox** is followed by 2057 different types. These types are a useful indication of the category of the object described in the page. Table 2 shows the top 24 **Infobox** categories with frequencies.

---

<sup>2</sup><http://www.wikipedia.org/>

<sup>3</sup><http://www.wiktionary.org/>

<sup>4</sup><http://en.wikiquote.org>

<sup>5</sup><http://download.wikimedia.org>

Figure 3: Exemple extractions from Taxobox

```

stevia|plant
squirrel|animal
springbok|animal
springbok antelope|animal
smelt|animal
smelts|animal
nitidulidae|animal
sap beetle|animal
scorpion|animal
streptococcus|eubacteria
snakes|animal

```

Table 2: Top 24 of Infobox types

|       |                   |      |                    |      |                  |
|-------|-------------------|------|--------------------|------|------------------|
| 47765 | Album             | 4857 | City               | 3379 | Biography        |
| 20042 | Film              | 4839 | actor              | 3163 | band             |
| 10056 | Single            | 4670 | Television episode | 2972 | Radio Station    |
| 9793  | musical artist    | 4353 | University         | 2936 | Software         |
| 9026  | CVG               | 3976 | Military Conflict  | 2789 | Australian Place |
| 7961  | Company           | 3925 | road               | 2681 | Swiss town       |
| 6307  | CityIT            | 3805 | Book               | 2540 | Broadcast        |
| 5070  | Indian urban area | 3449 | Mountain           | 2475 | Military Person  |

For all these extractions, we use the title of the page to complete the name on the box. This helps to capture variations. More variations can be found using the directive `#REDIRECT` that link a page directly to another page. By using the `title` tag, we build a file that contains terms variations. This variation is then used to enlarge the previous lists of terms.

Figure 4: Example of redirection extraction

```

torres del paine|cordillera del paine
whitpaines creek|wissahickon creek
cocker spainell|cocker spaniel
paine art center|paine art center and gardens
emma paine|the x factor uk series 1
torres del paine national park|cordillera del paine

```

### 3.3 Document Expansion with Extracted Knowledge

We have first proposed a basic indexing, using language model implemented into Lemur (see below run 09) with a MAP of 0.1342. We use corpus knowledge in the form of *pseudo-feedback* with 7 documents and 6 best terms. This gives an important improvement (run 21). We have test the different smoothing methods implemented in Lemur. The simple linear interpolation between document and corpus probability (Jelinek-Mercer method) gives good results with a weight of 0.8 between document and corpus statistics. The absolute discount seems to give similar results, a little better than the Bayesian dirichet one. With this setting we obtain a MAP of 0.1588. These results are similar with the CMU Toolkit (run 08). We have then expanded the documents with knowledge extracted from Wikipedia (run 15) with MAP of 0.1593. Using Wikipedia and pseudo feed back give similar results. Usually, it is the query which is expanded. We have chosen to

expand document, first because, in this collection, document are very short, but also because we wanted to generalize named entities to a more general category in order to answer queries with generic terms. Also, because we are using language model, hence statistics on document are important. When we expand all documents with the **Taxobox** information, and all **Geobox** and **Infobox** on the type **Mountain**, the increase is strong but similar to pseudo feedback.

We use little information from the images: the fact that they are black and white or colors using a tool based on maximum saturation (HVS) of the image. This information ("black and white photo") is automatically added to the description of all images (run 17). With this information from image we obtain 0.1707 of MAP, only by strongly enhancing the query 11. Finally the adding of more information with a manual built thesaurus (run 19), is a more efficient way to increase the MAP to 0.1806 from run 17. Combining this run with the visual one 04 give good results in run 20 (MAP 0.2295 but not as good as our other results. We can say that adding knowledge is effective, but not as good as one may expect. In fact run 11 with MAP of 0.2442 is better because it uses feedback information from the visual side. Hence it seems that inter-media feedback is more effective than adding external knowledge.

## 4 Description of Submitted Runs

Totally 27 runs were submitted including the CBIR runs, TBIR runs and mixed runs. To combine the ranking scores from different runs, the linear fusion method is utilized. The weight coefficients are empirically set. In the following we describe<sup>6</sup> the condition of each type of run. After the run number, the **V** means only visual, **T** only textual, **VrT** means pseudo relevance feedback on visual rank list using textual modality, **TrV** is also pseudo feedback but on text using visual ranking, **TiV** means a textual indexing but with some input from visual analysis, finally **TfV** means a late fusion between the two modalities. We give also the official MAP.

**01V\_HSV 0.0684** index at the HSV SVD space and cosine distance. It is the CBIR baseline with only one type of feature.

**02VrT\_HSV 0.0693** run 01 plus the PRF. The feedback documents are from the LM-based TBIR, i.e. run 08. It is provided to demonstrate the PRF effect. For the other 11 CBIR runs, the same method is carried out. Thus, 12 PRF-based CBIR runs are generated.

**03V\_12RUNS\_EQWEIGHT 0.1122** combine 12 CBIR runs using the equal weights, i.e. not tuning the weights.

**04V\_12RUNS\_WEIGHT 0.1204** combine 12 CBIR runs using the empirically tuned weights.

**05VrT\_12RUNS 0.1358** combine 12 PRF-based CBIR runs using the empirically tuned weights. Refer to run 02 for more details.

**06V\_FMLF\_HSVEDGEWW 0.0547** global feature matching (HSV histogram and EDGE histogram) with late fusion and weighting score for each feature. Normalize feature matrices (vectors) by Mean and Var. Linear fusion of rank lists with weighting scores 0.7 for HSV and 0.3 for edge.

**07V\_FMLF\_HSVEDGEGABORNW 0.0541** global feature matching (HSV histogram and EDGE histogram and Gabor features) with late fusion and equal weight for each feature.

**08T\_LM\_SWITTEN 0.1377** only LM-based TBIR system with CMU and Wittenbaum smoothing.

**09T\_LMKL\_S0M2D0.7 0.1342** only LM-based TBIR system with Lemur, no stemming. Smoothing used with interpolate smoothing strategy. The smoothing method is the "Absolute discount" with the default parameter 0.7 for the discount delta.

---

<sup>6</sup>to reduce the name size we remove the IPAL prefix of the run name

- 10T\_LSI 0.1322** only LSI based TBIR system.
- 11TrV\_LM\_12RUNSVISUAL 0.2442** LM-based TBIR, i.e., run 08, plus the PRF. The feedback documents are from the combined CBIR without PRF, i.e. run 04.
- 12TiV\_LM\_12RUNSVISUAL\_FB 0.2516** LM-based TBIR, i.e., run 08T, plus the PRF. The feedback documents are from the combined CBIR with PRF, i.e. run 05.
- 13TfV\_LM\_12RUNSVISUAL 0.2226** <sup>7</sup> combine runs 08 and 04.
- 14TfV\_LM\_FB\_12RUNSVISUAL 0.2833** combine runs 11 and 04.
- 15T\_WTm\_S0M2D0.8C6T6 0.1593** LM-based TBIR with document expansion using information automatically extracted from Wikipedia. It uses extracted plant, animals, geographic terms, mountains, and also information of the redirection. New terms are added to document to enlarge the matching. LM discount is adapted to 0.8 and top document in pseudo relevance feedback to 6 and keep 6 terms.
- 16TfV\_WTm\_12RUNSVISUAL 0.2180** combine runs 15 and 04.
- 17TiV\_WTm\_S0M2D0.8C6T6 0.1707** This run is an extension of run 15 using a black and white image detector based on HSV value of image. This information is injected into all corresponding texts. The matching is still done on text domain. This just boosts only one query, but boost globally the result.
- 18TiVfV\_T1.0V1.0 0.2165** The run 19 is fused with purely visual run 04. This late fusion is linear without weight, but with a linear transformation of all RSV output to the interval [1.0;0.00001] We do not choose to go to zero as minimum, to differentiate with document that are not retrieved, in either of the two modalities. We consider that a document not retrieved has a Relevance Status Value of zero.
- 19TiV\_WTmM\_S0M2D0.8C6T6 0.1806** This run is an extension of run 17, with a very small adds thesaurus manually extracted from Wikipedia. It is terms that are not from info boxes but are relevant to this collection.
- 20TiVfV\_T1.0V1.0 0.2295** The same simple fusion is performed than run 18 but with the textual run 19.
- 21T\_LMKL\_S0M2D0.7C9T6 0.1588** This run is a variation of run 09. Relevance feed back is used on text only, with 9 top documents used as feedback and 6 best terms added to the query. The improvement is important.
- 22V\_13RUNSVISUAL 0.1195** combine 12 CBIR runs and one run with the edge histogram using the equal weight, i.e. not tuning the weights.
- 23TfV\_LM\_FB\_13RUNSVISUAL 0.2732** combine runs 11 and 22.
- 24\_FMLF\_HSVEDGENW 0.0526** same as run 06 but using eual weight for fusion.
- 25V\_FMLF\_HSVEDGEGABORWW 0.0554** same as run 07. Linear fusion of two ranking lists is performed with weighting scores 0.6 for HSV and 0.3 for edge and 0.1 for gabor.
- 26V\_COOC\_PATCHES\_HSVEDGE\_1000CL 0.0369** index using Bag-of-vistern model with 1000 clusters.
- 27V\_COOC\_16PATCHES\_HSVEDGE\_1000CL\_KEYWORDFILTER 0.0311** same as run 26 but with filter of visual cluster by using top 100 most frequent keywords from text

---

<sup>7</sup>Due to some error this run has not being officially evaluated.



## 5 Results for Photographic Task

The best run is 14 with the MAP value 0.2833. It is ranked at the 6th place among 476 runs and at the 2nd place among automatic runs. This run is the mixed modality retrieval plus the CBIR based pseudo-relevance feedback (see section 2.4 (item 11) for the detailed description to the run). The best CBIR run reaches the MAP value 0.1204 for run 04 without the PRF. It is ranked at the 4th place for the visual based retrieval. Compared with the 1st run (INAOE/ INAOE-TIA-INAOE-VISUAL-EN-AN\_E, MAP: 0.1925) and the 2nd run (XRCE/ XRCE-IMG\_COMBFK, MAP: 0.1890), there are still room for us to improve our CBIR system. Now we see how PRF effects on the performance. First, for the CBIR system, the feedback from the TBIR seems have few effect on the performance. For example, with the PRF, the MAP of the HSV-based CBIR (run 02) is only increased to 0.0693 from 0.0684 (run 02). When combining 12 CBIR runs, the MAP is increased to 0.1358 (run 05) from 0.1204 (run 04). This should due to the diversity of visual content and divergence of visual features. However, the PRF has a significant improvement for the TBIR. For example, the MAP value of the LM-based TBIR is 0.1377 (run 08). When the CBIR based pseudo-relevance feedback (run 04) is applied, significant improvement is observed and its MAP value reaches 0.2442 (run 11).

## 6 Medical Task: Bayesian Network Conceptual Indexing

We propose a Bayesian network based model, inferred from [14], for capturing the concepts and semantic relationship between document concepts and query concepts. The conceptualisation is based on an external knowledge, named UMLS (Unified Medical Language System). The method of concept extraction is the same as our previous work in CLEF2006 [6], i.e. we use Metamap [1] for mapping concepts in English reports, and our tool for French and German reports after tagging by TreeTager. These concepts extracted are then used to build up the network for our Bayesian network based retrieval framework. Bayesian network is a Direct Acyclic Graph(DAG) [11], called also belief network, causal network. Our Bayesian network based model includes document nodes, query nodes and concept nodes and direct links between nodes.

There are two types of links: links which connect concepts nodes to documents or queries in which these concepts appear and links between document concept nodes and query concept nodes if there're semantic relations between them found in UMLS. These semantic relations are derived from UMLS relationship database files (MRREL.RRF for direct relation and MRHIER.RRF for hierarchical context of each instant of concept in different knowledge sources of UMLS). We decide the direction of links is on the path from documents to queries so that we can calculate the probabilistic inference from documents to queries in the network. Nodes which are pointed by a link are considered as child nodes.

The retrieval process is a probabilistic inference from document node to query node, considered as a 3-steps process:

- A document is observed which permits an initiation of prior probabilities of document concept nodes: in this model we chose tf.idf normalized as prior probability for document concept nodes.

$$P(c|D) = \frac{w_c}{\sqrt{\sum_{c' \in D} w_{c'}^2}}$$

with

$$w_c = tf * idf$$

- These prior probabilities of document concept nodes are then inferred to query concept nodes via semantic links between them by this formulas:

$$P(c|pa(c), D) = \max_i(\alpha * P(pa_i(c)|D))$$

in which  $pa(c)$  is the set of parent concepts of  $c$ ,  $\alpha$  is parameter used to estimate the uncertainty of the relatedness or "aboutness" between  $c$  and its parent according to the type of semantic link. We suppose  $\alpha$  may depend also on the quality of knowledge source who supply the relationship information. The bigger  $\alpha$  is, the more relevant the two concepts are.

We predefine  $\alpha$  empirically in the range  $[0, 1]$  according to the semantic relation corresponding or estimate it by the semantic similarity formulas of Leacock-Chorodo [7] in case of indirect hierarchical relations:

$$sim(c_1, c_2) = \log \frac{2 * L}{l(c_1, c_2)}$$

in which  $L$  is the maximum length of the taxonomy,  $l(c_1, c_2)$  is the minimum length of path from  $c_1$  to  $c_2$ .

- The relevance status value (RSV) of a query  $Q$  corresponding to a document  $D$  is the conditional probability of query node  $Q$  with the condition that document node  $D$  is observed. In this case we use the inference mechanism as [11] in which the probability of a node given their parents is calculated by link matrix. We chose weighted sum link matrix formulas in this case in other to take into account at the same time the inferred probability and weight of query concept nodes:

$$RSV(Q, D) = P(Q|pa(Q), D) = \frac{\sum_i p_i w_i}{\sum_i w_i}$$

$p_i$  is the probability of query concept node (in the set  $pa(Q)$ ) inferred from document concept nodes of  $D$ ;  $w_i$  is the weight of query concept node correspond to query  $Q$ . This weight is measured by *tf.idf* normalized similarly to document concept nodes.

The relevant status value in the ranking list then can be re-weighted by semantic groups of concepts: RSV is multiplied with the number of concepts in the matched concepts set that correspond to the three important semantic groups: Anatomy, Pathology, and Modality. The purpose is to emphasize the important of these three semantic groups considering the structure of queries.

## 6.1 Results of Medical Runs

For the CLEF image medical 2007 collection [10], we analyse the new image sets(MyPacs and Endoscopic) the same way as the collection last year.

In the submitted runs, based on predefined set of semantic relations in UMLS, we use isa-indirect relations in four runs and different types of relations in the two other runs: isa, parent-child other than isa(PAR-CHD), broader-narrower(BR-RN), similar or alike(RL), related or possibly synonym(RQ). The value of  $\alpha$  in all of these submitted runs is predefined in the range  $(0, 1)$ , corresponding: 0.1, 0.2, 0.3, 0.4. In addition to these runs, we experiment another run which replaces predefined  $\alpha$  by semantic similarity measurement of Leacock-Chorodo for isa-indirect relation. This method yeilds similar result. Results of our submitted runs are reported in table 3.

## 7 Conclusion

The Bayesian model proposed here, subsume the VSM model with cosine similarity, and in addition take into account the semantic relationship between documents concepts and query concepts in an unified framework. Experimentation shows that this model can enhance the VSM by adding the semantic relatedness between concepts. Improvements on relationship weighting issue as well as performance of model are our further study.

Table 3: MAP results of CLEFMed2007

| Run                  | Isa | PAR-CHD | BR-RN | RL    | RQ    | Map    | R-prec |
|----------------------|-----|---------|-------|-------|-------|--------|--------|
| IPAL-TXT-BAY-ISA0.1  | 0.1 | 0       | 0     | 0     | 0     | 0.3057 | 0.332  |
| IPAL-TXT-BAY-ALLREL2 | 0.2 | 0.01    | 0.01  | 0     | 0     | 0.3042 | 0.333  |
| IPAL-TXT-BAY-ALLREL1 | 0.2 | 0.01    | 0.01  | 0.001 | 0.001 | 0.304  | 0.3338 |
| IPAL-TXT-BAY-ISA0.2  | 0.2 | 0       | 0     | 0     | 0     | 0.3039 | 0.3255 |
| IPAL-TXT-BAY-ISA0.3  | 0.3 | 0       | 0     | 0     | 0     | 0.2996 | 0.3212 |
| IPAL-TXT-BAY-ISA0.4  | 0.4 | 0       | 0     | 0     | 0     | 0.2935 | 0.3177 |

For the photographic task, we have improved our CBIR system the use of multiple visual features and different indexing techniques. Combined with the cross-modality pseudo-relevance feedback and text based search, our system obtains a good performance. From our results, we find 1) CBIR is an important component to search multi-modality image database, especially when the text description has weak discrimination; 2) cross-modality pseudo-relevance feedback can enhance the performance. But it has much more effect on the text based search than on the visual based search. The use of external knowledge like Wikipedia gives an extra bonus to the results but is less effective than expected. The large size of this set tends also to produce confusion because of lack of disambiguation. This disambiguation problem occurs mostly on small terms.

## References

- [1] Alan R. Aronson. Metamap: Mapping text to the umls metathesaurus. <http://mmtx.nlm.nih.gov/docs.shtml>, July 2006.
- [2] Ricardo A. Baeza-Yates and Berthier A. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press / Addison-Wesley, 1999.
- [3] Jerome R. Bellegarda. Exploiting latent semantic information in statistical language modeling. *Proceedings of the IEEE*, 88(8):1279–1296, August 2000.
- [4] Sheng Gao, Joo-Hwee Lim, and Qibin Sun. An integrated statistical model for multimedia evidence combination. In *ACM Multimedia'07*, 2007.
- [5] Sheng Gao, Wen Wu, Chin-Hui Lee, and Tat-Seng Chua. A maximal figure-of-merit (mfom)-learning approach to robust classifier design for text categorization. *ACM Trans. Inf. Syst.*, 24(2):190–218, 2006.
- [6] Caroline Lacoste, Jean-Pierre Chevallet, Joo-Hwee Lim, Xiong Wei, Daniel Raccoceanu, Diem Le Thi Hoang, Roxana Teodorescu, and Nicolas Vuillenemot. Ipal knowledge-based medical image retrieval in imageclefmed 2006. In *Working Notes for the CLEF 2006 Workshop, 20-22 September, Medical Image Track, Alicante, Spain*, 2006.
- [7] C. Leacock and M. Chodorow. Combining local context and WordNet similarity for word sense identification. In C. Fellbaum, editor, *WordNet: An Electronic Lexical Database*, chapter 11, pages 265–284. MIT Press, 1998.
- [8] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, 2004.
- [9] Nicolas Maillot, Jean-Pierre Chevallet, Vlad Valea, and Joo Hwee Lim. Ipal inter-media pseudo-relevance feedback approach to imageclef 2006 photo retrieval. In *Working Notes for the CLEF 2006 Workshop, 20-22 September, Alicante, Spain*, 2006.

- [10] Henning Müller, Thomas Deselaers, Eugene Kim, Jayashree Kalpathy-Cramer, Thomas M. Deserno, Paul Clough, and William Hersh. Overview of the ImageCLEFmed 2007 medical retrieval and annotation tasks. In *Working Notes of the 2007 CLEF Workshop*, Budapest, Hungary, sept 2007.
- [11] Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988.
- [12] Jay M. Ponte and W. Bruce Croft. A language modeling approach to information retrieval. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 275–281, New York, NY, USA, 1998. ACM Press.
- [13] Clarkson P. R. and Rosenfeld R. Statistical language modeling using the cmu-cambridge toolkit. In *Eurospeech 1997*, 1997.
- [14] Howard Turtle and W. Bruce Croft. Evaluation of an inference network-based retrieval model. *ACM Trans. Inf. Syst.*, 9(3):187–222, 1991.