DOKUZ EYLÜL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

# AN EXPANSION AND RERANKING METHOD FOR ANNOTATION BASED IMAGE RETRIEVAL FROM WEB

by

Deniz KILINÇ

October, 2010
İZMİR

# AN EXPANSION AND RERANKING METHOD FOR ANNOTATION BASED IMAGE RETRIEVAL FROM WEB

**A Thesis Submitted to the**

**Graduate School of Natural and Applied Sciences of Dokuz Eylül University**

**In Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Computer Engineering, Computer Engineering Program**

**by**

**Deniz KILINÇ**

**October, 2010**

**İZMİR**

## Ph.D. THESIS EXAMINATION RESULT FORM

We have read the thesis entitled **"AN EXPANSION AND RERANKING METHOD FOR ANNOTATION BASED IMAGE RETRIEVAL FROM WEB"** completed by **DENİZ KILINÇ** under supervision of **PROF. DR. ALP R. KUT** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Doctor of Philosophy.

Prof. Dr. Alp R. KUT

Supervisor

| | |
|---|---|
| Prof. Dr. Yalçın ÇEBİ | Prof. Dr. Ender YAZGAN BULGUN |
| Thesis Committee Member | Thesis Committee Member |

| | |
|---|---|
| Asst. Prof. Dr. Adil ALPKOÇAK | Assoc. Prof. Dr. Onur DEMİRÖRS |
| Examining Committee Member | Examining Committee Member |

Prof.Dr. Mustafa SABUNCU

Director
Graduate School of Natural and Applied Sciences

# ACKNOWLEDGMENTS

I would like to show my gratitude to my supervisor, Prof. Dr. Alp R. KUT and to my advisor, Asst. Prof Adil ALPKOÇAK for their guidance and encouragement. It was an honor for me to work with them. I am heartily thankful to Asst. Prof Adil ALPKOÇAK, whose support from the initial to the final level enabled me to develop an understanding of the subject. I appreciate his knowledge and his assistance in reviewing and guiding me on my reports and papers.

I would like to thank the members of thesis committee for their comments and suggestions.

Finally, I would like to thank to my wife Sibel KILINÇ, my daughter Ada KILINÇ, my parents Kayber KILINÇ, Hadice KILINÇ, Türkan KOPARAN and Mehmet KOPARAN. This thesis would not have been possible without their support and love completing.

Deniz KILINÇ

# AN EXPANSION AND RERANKING METHOD FOR ANNOTATION BASED IMAGE RETRIEVAL FROM WEB

## ABSTRACT

Current state-of-the-art in image retrieval has two major approaches: content-based image retrieval (CBIR) and annotation based image retrieval (ABIR). Annotation-based image retrieval (ABIR) simply uses text retrieval techniques on annotations generally done by human.

In this thesis, we propose a new expansion and reranking approach for annotation based image retrieval (ABIR) from Web images. Our suggestion considers an image retrieval system using the surrounding texts nearby the image in a web page as annotations. However, annotations may include too much and uninformative text such as copyright notice, date, author etc. In order to choose indexing terms effectively, we propose a term selection approach, which first expands the document using WordNet, and then selects descriptive terms among them. Notably, we applied this term selection methodology to both document and query. This is because applying either of documents or query does not help to increase retrieval performance. On the other hand, documents and queries become more exhaustive than original. Consequently, this results high recall with low precision in retrieval. Thus, we also proposed a two-level reranking approach. Experiments have demonstrated that that document expansion and reranking plays an important role in text-based image retrieval and two-level reranking betterments the retrieved results by increasing precision.

**Keywords**: Information Retrieval, Image Retrieval, Query Expansion, Document Expansion, Reranking, WordNet

# WEBDEN BETİMLEME TABANLI GÖRÜNTÜ ERİŞİMİ İÇİN GENİŞLETME VE TEKRAR SIRALAMA YÖNTEMİ

## ÖZ

Son gelinen noktada, resimlere erişim tekniği olarak görülen iki önemli yaklaşım bulunmaktadır: İçeriğe dayalı bilgi erişimi (İDBE) ve betimlemeye dayalı bilgi erişimi (BDBE). BDBE, resimlere insanlar tarafından iliştirilen notları, metin tabanlı arama yöntemi ile bulmaya dayanır.

Bu tezde, BDBE tekniği ile webe yüklenmiş resimlerin erişim işlemini geliştiren yeni *"genişletme"* ve *"tekrar sıralama"* yaklaşımları sunulmuştur. Çalışma, resimler web ortamına yüklenirken onlara iliştirilen notlar kullanılarak bir resim sorgulama ve erişim sistemi tasarlanması üzerine kurulmuştur. Resimlere iliştirilen bu notların; tarih, yükleyici adı ve kullanım hakları gibi sorgulama açısından gereksiz bilgileri içerdiği unutulmamalıdır. Önerilen sistemde, en efektif terimleri seçmek için WordNet kullanılarak resim dokümanları genişletilmiş ve sonrasında anlamlı terimleri seçilmiştir. Resim erişim performansını arttırmak için kullanılan genişletme tekniği hem sorgular hem de dokümanlar üzerinde aynı şekilde uygulanmıştır. Yapılan bu genişletme işlemleri daha genişlemiş doküman ve sorgular oluşturduğu için anma (recall) seviyesini arttırırken, hassasiyet (precision) seviyesini düşürmüştür. Düşen hassasiyet seviyesini arttırmak için iki seviyeli tekrar sıralama yaklaşımı sunulmuştur. Yaptığımız deneyler, önerdiğimiz genişletme ve tekrar sıralama yaklaşımlarının web ortamına yüklenmiş resimlerin metin tabanlı aranması işleminin iyileştirilmesinde önemli bir rol oynayabileceğini ispatlamıştır.

**Anahtar sözcükler**: Bilgi Erişimi, Resim Erişimi, Sorgu Genişletme, Belge Genişletme, Tekrar Sıralama, WordNet

## CONTENTS

# CHAPTER ONE
# INTRODUCTION

## 1.1 Background

In recent years, there has been a tremendous increase of available image data in both scientific and consumer domains, as a result of current rapid advances of Internet and multimedia technology. As hardware has improved and available bandwidth has grown, the size of digital image collections has reached terabytes and this amount constantly grows day by day. The importance of this image information depends on how easily we can search, retrieve and access to it.

Current state-of-the-art in image retrieval has two major approaches: content-based image retrieval (CBIR) and annotation based image retrieval (ABIR). A basic difference between ABIR and CBIR is related to the values of textual and visual information in image retrieval.

As can be seen in many of today's image retrieval systems, such as web search engines and clip-art searching software, ABIR is considered practical in many general settings. Two user studies suggest the importance of textual information (ABIR) in image retrieval. Hughes et al. (2003) revealed that users of video retrieval systems tend to use textual information more often than visual information to validate their search results. Another study found a similar result for photo images (Choi & Rasmussen, 2002). Consequently, textual information should play a central role in visual information retrieval. CBIR approach works only on the images by extraction of visual primitives like color, texture or shape, which is computationally expensive and become quite infeasible as image collection gets larger.

Annotation-based image retrieval (ABIR) simply uses text retrieval techniques on textual annotations of images which are generally done by human. In World Wide Web environment, much of the image content is insufficiently supplied with textual metadata and, it is not realistic to expect annotation of such huge number of images by hand. A simple alternative is to use information, in the form of textual data, around the image such as, image file-name, html tags.

Notably, the text surrounding images might be more descriptive and is usually includes descriptions implicitly made by page designer. All these textual data could be stored with the image itself, and could be used as annotation of images associated with unstructured metadata. In fact, the surrounding textual content should be considered since it is probable that surrounding text includes some form of human generated descriptions of the images, which is somehow *closer semantic* interpretation.

## 1.2 Problem Definition

It is hard to extract low-level features from web images or manually annotate them. Many techniques for extracting of low level cues are distinguished by the characteristics of domain-images. But web images are heterogeneous collection of images that are searched for by users with diverse information needs. So these techniques won't be proper. Also, performance of these techniques is challenged by various factors like image resolution, intra-image illumination variations, non-homogeneity of intra-region and inter-region textures, multiple and occluded objects etc.

The other major difficulty, described as *semantic gap problem* of CBIR systems in the literature, is a gap between inferred understanding / semantics by pixel domain processing using low level cues and human perceptions of visual cues of given image. In other words, there exists a gap between mapping of extracted features and human perceived semantics. The dimensionality of the difficulty becomes adverse because of subjectivity in the visually perceived semantics, making image content description a subjective phenomenon of human perception, characterized by human psychology, emotions, and imaginations. Furthermore, the user query must be entered in the form of that modality with low level image features, which is not simple to do.

ABIR can be an applicable approach for image retrieval for Web images like Wikipedia, when surrounding text is used as annotation which is implicit descriptions of the images and is closer semantic interpretation. However it requires new expansion and reranking techniques to improve its retrieval performance results.

## 1.3 Goal of Thesis

The goal of the thesis is to propose a new expansion and reranking method for ABIR from web resources like Wikipedia images. To increase recall, both documents and queries are expanded in expansion step using same methods. Although expansion step increases the recall, at the same time, it decreases the precision. Consequently, a two-level reranking method is applied to increase precision.

## 1.4 Methodology

We propose a new expansion technique which expands the queries through local analysis which is one of the most effective methods of reformulating queries without relying on user input. We use WordNet (Miller, 1990) online lexical system, Word Sense Disambiguation (WSD) technique and WordNet similarity functions. On the other hand, if it is true for queries, documents (i.e., image annotations) must be expanded, as well as queries. Text retrieval community studied query expansion extensively. However, in literature, document expansion has not been thoroughly researched for information retrieval. From the past research whether the document expansion can improve the retrieval effectiveness or how to improve is not obvious (Singhal & Pereira, 1999; Billerbeck & Zobel, 2005; Ide & Salton, 1971; Li & Meng, 2003).

Since, document and query expansion generally result *high recall* with *low precision* in retrieval, a two-level reranking method is introduced to increase precision by reordering the result sets. The first level forms a narrowing-down operation and includes re-indexing. This novel method is based on filtering out non-relevant documents and reducing both the number of documents and the number of terms. It shrinks down the initial VSM data into more manageable size so that we perform more complex cover coefficient (CC) based reranking algorithm upon in the second level.

We evaluated the proposed *ABIR strategy with expansion and reranking method* on ImageCLEF's WikipediaMM task (Tsikrika & Kludas, 2009) which provides a test bed for the system-oriented evaluation of Wikipedia images by using both WikipediaMM 2008 and WikipediaMM 2009 topics/queries. Evaluation results show that ABIR approach is promising for current state-of-the-art for image retrieval.

## 1.5 Contribution of Thesis

The main contribution of this thesis is to propose an ABIR system using new expansion technique for both documents and queries (WordNet, WSD, similarity functions) and applying two-level reranking approaches to increase precision.

## 1.6 Thesis Organization

In this chapter, we have stated what we are trying to accomplish, what is our goal and methodology, and our contribution to the field. The rest of the thesis is organized as follows. The next, Chapter 2 presents a literature survey on CBIR, ABIR including expansion (document, query) and reranking methods. We also express VSM (Vector Space Model), term-weighting and normalization in Chapter 2.

In Chapter 3, we present and sample our expansion method that is developed for annotation based image retrieval (ABIR) for both documents and queries using WordNet, WSD and Similarity functions. Chapter 4 presents proposed new reranking method which includes two-level: The first level forms a narrowing-down phase of search space while second level includes a cover coefficient based reranking. In chapter 5, proposed system's experimentation results on the ImageCLEF2009 WikipediaMM task are showed. The results we obtained are superior to any participating approaches and our approach has obtained the best four ranks, in text-only image retrieval. The results also showed that document expansion and reranking plays an important role in ABIR. The last chapter concludes the thesis by providing the results we obtained and offers future works on this topic.

# CHAPTER TWO
# DEFINITIONS AND RELATED WORKS

## 2.1 Introduction

This chapter starts with the history overview of image retrieval. In section 2.3 and 2.4, definitions and related works for Image Retrieval, CBIR and ABIR are presented. Preliminary definitions are also explained for VSM (Vector Space Model), term weighting and document normalization (cosine, pivoted unique).

In section 2.8 and 2.9 expansion and reranking methods' definitions with related researches for image and document retrieval are described. Most of the researches in annotation based image retrieval are limited to use of one method, which is usually query expansion or relevance feedback or reranking. Combined use of both expansion (i.e., document and query) and reranking method is unique for our research.

## 2.2 History of ABIR and CBIR

ABIR approaches were first experimented for image retrieval, where textual annotations were manually added to each images and the retrieval process was performed using standard database management systems (Chang & Fu, 1980; Chang & Kunii, 1981; Chang & Fu, 1979). In the early 90's, with the growth of image collections, manually annotation approach of the images became inoperable.

As a result, CBIR was proposed which is based on extracting low-level visual content such as color, texture, or shape. The extracted features are then stored in a database and compared to an example image query.

Many studies based on content based retrieval, differs on the techniques used for extracting and storing features (El Kwae & Kabuka, 2000; Ogle & Stonebraker, 1995; Wu J. K., 1997) and on the image searching methods (Flickner, et al., 1995;

Santini & Jain, 2000). With the expansion of World Wide Web, image retrieval interest has veered (Frankel, Swain, & Athitsos, 1996; Lew, Lempinen, & Huijsmans, 1997).

In the Web, the images are usually stored with image file-name, html tags and surrounding text. In the course of time, multi-modal systems have been suggested to improve image search results by using the combination of textual information with image feature information (Wu, Iyengar, & Zhu; Wong & Yao, 1995).

### 2.3 CBIR (Content Based Image Retrieval)

CBIR is the science of how we can index and retrieve images based on its low level features. When we talk about low-level features, it means, the most basic features of an image. Images can typically be divided into three different low-level features: Color, shape and texture.

- *Color;* Each pixel in a digital image consists of a color element. In a grey-scale image this color element typically range from 0 to 255, where 0 is black, 255 is white, and the values between is the different shades of grey from black to white. In a color image, let's say with 24 bits color resolution, (which means that 24 bits are used for color information for each pixel), an (often even) piece of the 24 bits is assigned to each of the three color components in the image.

  The most common color space used is RGB. Color Space is defined as a model for representing color in terms of intensity values RGB stands for Red-Green-Blue, and in this example the 24 bit color image, 8 bits is used to represent each of the components. In this example the color components makes it possible to represent (2^8)^3 different colors, which is more than the human eye can differentiate from. Even though RGB is the most common color space used, it is not always the best choice when working with CBIR.

- *Shape* is the contours and shapes of objects represented in the image. The process of extracting shapes often goes like this: First the contours in the image are found, and then we segment the image into the different contours and index these.

  Finding contours can be obtained by using chain codes which is an algorithm that "walks" around the edge of regions, creating a set of straight lines around the region. A region could for instance be an area of similar color or an area that is in some way different from the rest of the image. These lines can be further simplified with polygon approximation which makes the lines less jagged.

- *Texture* is a way of extracting what kind of "surface" an image or object has. Different features which describes texture. *Contrast* feature measured by using four parameters: dynamic range of grey levels of the image, polarization of the distribution of black and white on grey-level histograms or ratio of black and white areas, sharpness of edges and period of repeating patterns. *Directionality* measures elements shape and placement. *Line likeness* measures are concerned with the shape of a texture element. *Regularity* measures variation of an element placement rule. *Roughness* measures whether or not an object is smooth. A polished ball will have little roughness, while a mountain will have a rough surface (at least close up).

### *2.3.1 Some CBIR Applications*

IBM's *QBIC* (Query by Image Content) system (Flickner et al, 1995) is probably the best-known of all image content retrieval systems. It is available commercially either in standalone form, or as part of other IBM products such as the DB2 Digital Library. It offers retrieval by any combination of color, texture or shape – as well as by text keyword. Image queries can be formulated by selection from a palette, specifying an example query image, or sketching a desired shape on the screen.

The system extracts and stores color, shape and texture features from each image added to the database, and uses R*-tree indexes to improve search efficiency. At search time, the system matches appropriate features from query and stored images, calculates a similarity score between the query and each stored image examined, and displays the most similar images on the screen as thumbnails. The newer versions of the system incorporates more efficient indexing techniques, an improved user interface, the ability to search grey-level images, and a video storyboarding facility.

*Blobworld* (Carson et al, 1999) is a CBIR system developed at University of California, Berkeley. The system automatically extracts the regions of an image, which roughly correspond to object or parts of objects. It allows users to query for images based on the objects they contain. The user first selects a category, which already limits the search space. In an initial image, the user selects a region (blob), and indicates the importance of the blob. Next, the user indicates the importance of the blob's color, texture, location, and shape. More than one region can be used for querying. Their approach is useful in finding specific objects and not, as they put it, "stuff" as most systems which concentrate only on "low level" features with little regard for the spatial organization of those features. It allows for both textual and content-based searching.

*Simplicity (Semantics sensitive Integrated Matching for Picture Libraries)* (Wang, Li., & Wiederhold, 2001) is an image retrieval system, which uses a wavelet-based approach for feature extraction, semantics classification methods, and integrated region matching based upon image segmentation. Their system classifies images into semantic categories such as textured-nontextured, graph photograph.

Potentially, the categorization enhances retrieval by permitting semantically adaptive searching methods and narrowing down the searching range in a database. A measure for the overall similarity between images is developed using a region-matching scheme that integrates properties of all the regions in the images. For the purpose of searching images, they have developed a series of statistical image classification methods.

*WebSeek* (Smith, 1997) collects its content by a collection processes through Web robots, though it has the advantage of video search and collection as well. It was developed at Columbia University. WebSeek makes text-based and color based queries through a catalogue of images and videos.

Color is represented by means of a normalized 166-bin histogram in the HSV color space. For the query, user initiates a query by choosing a subject from the available catalogue or entering a topic. The results of the query may be used for a color query in the whole catalogue or for sorting the result list by decreasing color similarity to the selected item. Also, the user has the possibility of manually modifying an image/video color histogram before reiterating the search.

## 2.4 ABIR (Annotation Based Image Retrieval)

CBIR is suitable for "find-similar" tasks, in which, searched images may not differ significantly in their appearances, and so the facile similarities of the images are more important than the semantic contents. Examples are medical diagnoses based on the comparison of X-ray pictures with past cases, and for finding the faces of criminals from video shots of a crowd (crime prevention).

Applications that involve more semantic relationships cannot be dealt with by CBIR, even if extensive image processing procedures are applied. For instance, in the collecting of the photos regarding the "tennis player", it is difficult what kind of images should be used for the querying. This is simply because visual features cannot fully represent concepts. Only texts or words can do that.

Annotation-based image retrieval (ABIR) is a kind of text based retrieval system that uses textual annotations of images which are generally done by human. A basic difference between ABIR and CBIR is related to the values of textual and visual information in image retrieval.

### *2.4.1 Sparseness and Annotation Quality in ABIR*

Term co-occurrence frequencies, is often sparse in IR. In annotated images, the occurrences of words are especially limited because they must be assigned only for indexing purposes and the need for such extra effort is not appreciated. The worst annotation may be only one word, which is the file name of the image. Handling such severe word sparseness is one important research topic in ABIR.

The problem of word sparseness may be mitigated by incorporating external knowledge such as thesauri, like WordNet (Miller, 1990) that explicitly identify the relationships between words. This approach is frequently studied in textual IRs and is applicable to ABIR as well.

In addition to explicit knowledge, implicit information can be utilized in ABIR. Zhou (2002) suggested that CBIR is limited because it relies solely on low-level visual features. They proposed the use of textual information within the CBIR framework. They also mentioned the problem of word sparseness. They used relevance-feedback (RF) for estimating word associations in annotated images. RF can be considered contextual information at the user system interaction level.

*Quality of the annotations* should be taken into account, when retrieving images based on annotations. We assume that manually assigned annotations are usually more reliable than automatically assigned ones. Because of the cost, however, annotations are sometimes assigned automatically. Two types of methods are frequently used to assign textual information to images.

One method is based on information extraction techniques. For example, some textual information corresponding to images on the WWW can be extracted from their surrounding texts or anchor texts linked to the images. If the extraction rules are carefully designed, the acquired annotations may be relevant to the images. However, because there are usually exceptions in the data that are not consistent with the assumptions, the extracted annotations may contain noise.

The other method is based on classification techniques. The development of procedures for assigning keywords to a given image is an active research topic (e.g., Jeon et al (2003)). Such automatic annotation can be regarded as a type of multi-class image classification. Although classification itself has been relatively well studied, automatic annotation cannot be performed easily.

## 2.5 Query Formulation for Image Retrieval

In a typical usage of image retrieval system, user should be able to search an image database for images that express the desired information or (s)he may process an image and (s)he is interested in and wants to find images from the database that are similar to the query image. Different implementations of image retrieval make use of different types of user queries.

- *Query by example,* the user searches with a query image (supplied by the user or chosen from a random set), and the software finds images similar to it based on various low-level criteria.

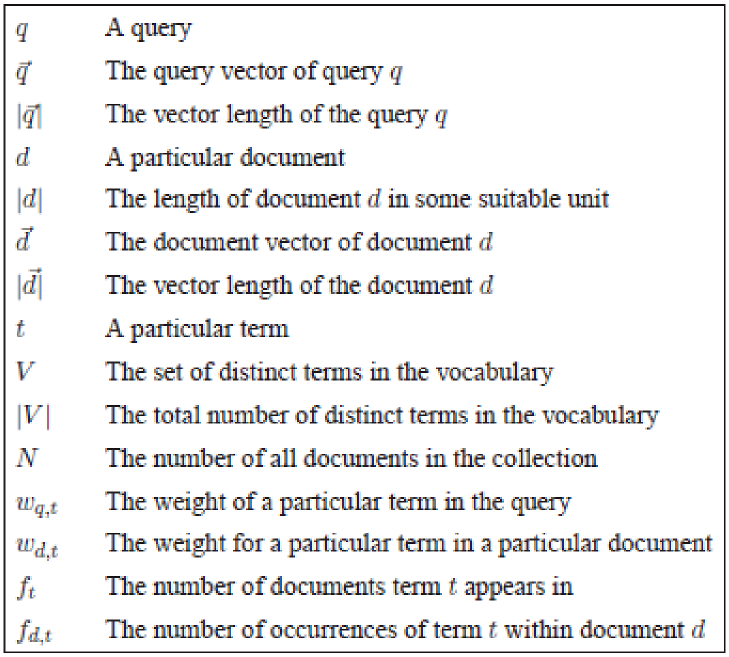
- *Query by keyword,* the user submits a keyword and software locates images that are related with that keyword. Traditional systems use this approach and retrieve the results by exact match with annotations. For content based image retrieval systems, query operation does not performed on manual annotations instead system makes search on annotations that are estimated automatically (auto-annotation).

- *Query by sketch,* user draws a rough approximation of the image they need and for example with colored regions and the system locates images whose layout matches the sketch.

- Other methods include specifying the proportions of colors desired and searching for images that contain an object given in a query image.

## 2.6 VSM for Image Retrieval

Vector Space Model (VSM) is widely used in information retrieval where each document is represented as a vector, and each dimension corresponds to a separate term. If a term occurs in the document then its value in the vector is non-zero. Model employs a ranking algorithm that tries to rank documents in order of how much of an overlap there is between the terminology of the query and each document (Salton, 1971; Bookstein, 1982), where relatively rare terms have a comparatively high weight. All queries and documents are represented as vectors in |V| dimensional space, where V is the set of all distinct terms in the dataset. Basically, documents are ranked by the magnitude of the angle between the document vector and the query vector. VSM notation is summarized in Figure 2.1.



| | |
|---|---|
| $q$ | A query |
| $\vec{q}$ | The query vector of query $q$ |
| $\lvert\vec{q}\rvert$ | The vector length of the query $q$ |
| $d$ | A particular document |
| $\lvert d\rvert$ | The length of document $d$ in some suitable unit |
| $\vec{d}$ | The document vector of document $d$ |
| $\lvert\vec{d}\rvert$ | The vector length of the document $d$ |
| $t$ | A particular term |
| $V$ | The set of distinct terms in the vocabulary |
| $\lvert V\rvert$ | The total number of distinct terms in the vocabulary |
| $N$ | The number of all documents in the collection |
| $w_{q,t}$ | The weight of a particular term in the query |
| $w_{d,t}$ | The weight for a particular term in a particular document |
| $f_t$ | The number of documents term $t$ appears in |
| $f_{d,t}$ | The number of occurrences of term $t$ within document $d$ |

Figure 2.1 Summary of VSM notations.

## 2.7 Term Weighting and Normalization

In VSM, term weighting is an important aspect of modern text retrieval systems. There are three major parts that affects the importance of a term in a text, which are the term frequency factor ($tf$), the inverse document frequency factor ($idf$), and document length normalization. Cosine normalization is the mostly used normalization technique in the vector space model. Normalization factor is computed as in the Equation 2.1.

$$\sqrt{w_1^2 + w_2^2 + \ldots + w_t^2} \tag{2.1}$$

where each $w$ equals $(tf \times idf)$ as in the equation 2.2.

$$w_{ki} = \frac{tf_{ik} \log (N/n_k)}{\sqrt{\sum_{\mathrm{k}=1}^{\mathrm{n}} (tf_{ik})^2 \, [\log (N/n_k)]^2}} \tag{2.2}$$

where $t_k$ is the $k^{th}$ term in document $d_i$. $tf_{ik}$ is the frequency of word $t_k$ in document $d_i$. $\log(\mathrm{N}/\mathrm{n_k})$ is inverse document frequency of word $t_k$ in dataset. $n_k$ is the number of documents containing the word $t_k$. *N* is the total number of document in dataset.

Table 2.1 presents a VSM example (Grossman & Frieder, 2004). In the example, we suppose that we search an IR system for the query "gold silver truck". The database collection consists of three documents (D = 3) with the following content. Retrieval results are summarized in the following table;

D1: "Shipment of gold damaged in a fire"
D2: "Delivery of silver arrived in a silver truck"
D3: "Shipment of gold arrived in a truck"

Table 2.1 VSM example, documents and query

| **Q: “gold silver truck”**<br>**D1: “Shipment of gold damaged in a fire”**<br>**D2: “Delivery of silver arrived in a silver truck”**<br>**D3: “Shipment of gold arrived in a truck”**<br>**D=3; IDF=log(D/df$_i$)** | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **Counts of tf$_i$** | | | | | | **Weights: w$_i$=tf$_i$*IDF$_i$** | | | |
| **Terms** | **Q** | **D$_1$** | **D$_2$** | **D$_3$** | **df$_i$** | **D/df$_i$** | **IDF$_i$** | **Q** | **D$_1$** | **D$_2$** | **D$_3$** |
| A | 0 | 1 | 1 | 1 | 3 | 3/3=1 | 0 | 0 | 0 | 0 | 0 |
| Arrived | 0 | 0 | 1 | 1 | 2 | 3/2=1.5 | 0.1761 | 0 | 0 | 0.1761 | 0.1761 |
| damaged | 0 | 1 | 0 | 0 | 1 | 3/1=3 | 0.4771 | 0 | 0.4771 | 0 | 0 |
| Delivery | 0 | 0 | 1 | 0 | 1 | 3/1=3 | 0.4771 | 0 | 0 | 0.4771 | 0 |
| Fire | 0 | 1 | 0 | 0 | 1 | 3/1=3 | 0.4771 | 0 | 0.4771 | 0 | 0 |
| Gold | 1 | 1 | 0 | 1 | 2 | 3/2=1.5 | 0.1761 | 0.1761 | 0.1761 | 0 | 0.1761 |
| İn | 0 | 1 | 1 | 1 | 3 | 3/3=1 | 0 | 0 | 0 | 0 | 0 |
| Of | 0 | 1 | 1 | 1 | 3 | 3/3=1 | 0 | 0 | 0 | 0 | 0 |
| Silver | 1 | 0 | 2 | 0 | 1 | 3/1=3 | 0.4771 | 0 | 0 | 0.9542 | 0 |
| shipment | 0 | 1 | 0 | 1 | 2 | 3/2=1.5 | 0.1761 | 0.1761 | 0.1761 | 0 | 0.1761 |
| Truck | 1 | 0 | 1 | 1 | 2 | 3/2=1.5 | 0.1761 | 0 | 0 | 0.1761 | 0.1761 |

Column definitions of VSM example are followed as;

- Columns 1 - 5: First, we construct an index of terms from the documents and determine the term counts *tfi* for the query and each document *Dj.*
- Columns 6 - 8: Second, we compute the document frequency *di* for each document. Since IDFi = log(*D*/*dfi*) and *D* = 3, this calculation is straightforward.
- Columns 9 - 12: Third, we take the *tf***IDF* products and compute the term weights. These columns can be viewed as a sparse matrix in which most entries are zero.

Table 2.2 VSM example, similarity analysis

| $|D_i| = \sqrt{\sum_i w_{i,j}{}^2}$ |
|---|
| $|D_1| = \sqrt{0.4771^2 + 0.4771^2 + 0.1761^2 + 0.1761^2} = \sqrt{0.5173} = 0.7192$ |
| $|D_2| = \sqrt{0.1761^2 + 0.4771^2 + 0.9542^2 + 0.1761^2} = \sqrt{1.2001} = 1.0955$ |
| $|D_3| = \sqrt{0.1761^2 + 0.1761^2 + 0.1761^2 + 0.1761^2} = \sqrt{0.1240} = 0.3522$ |
| $|D_i| = \sqrt{\sum_i w_{Q,j}{}^2}$ |
| $|Q| = \sqrt{0.1761^2 + 0.4771^2 + 0.1761^2} = \sqrt{0.2896} = 0.5382$ |

Similarity analysis starts with the computation of all vector lengths (zero terms ignored) for each document and query as presented in Table 2.2. Next, we compute all dot products (zero products ignored) as presented in Table 2.3.

Table 2.3 VSM example, dot products

| $Q\bullet D_i = \sum_i w_{Q,j}\, w_{i,j}$ |
|---|
| $Q\bullet D_1$=0.1761*0.1761=0.0310 |
| $Q\bullet D_2 = 0.4771 * 0.9542 + 0.1761 * 0.1761 = 0.4862$ |
| $Q\bullet D_3 = 0.1761 * 0.1761 + 0.1761 * 0.1761 = 0.0620$ |

Now we calculate the similarity values as presented in Table 2.4.

Table 2.4 VSM example, similarity calculation

| $Sim(Q, D_i) = \frac{Q\bullet D_i}{\lvert Q\rvert * \lvert D_i\rvert}$ |
|---|
| $Sim(Q, D_1) = \frac{0.0310}{0.5382 * 0.7192} = 0.0801$ |
| $Sim(Q, D_2) = \frac{0.4862}{0.5382 * 1.0955} = 0.08246$ |
| $Sim(Q, D_3) = \frac{0.0620}{0.5382 * 0.3522} = 0.3271$ |

Table 2.5 presents final ranked documents in descending order according to the similarity values.

Table 2.5 VSM example, final ranking

| **Rank** | **Doc** | **Rank Score** |
|---|---|---|
| 1 | D2 | 0.8246 |
| 2 | D3 | 0.3271 |
| 3 | D1 | 0.0801 |

### *2.7.1 Pivoted Unique Normalization*

Since the lengths of the document vectors are converted into unit vectors, the information content is deformed for longer documents, which contain more terms with higher $tf$ values and also more distinct terms in cosine normalization.

Pivoted Unique Normalization is a modified version of the classical cosine normalization and $(tf \times idf)$. A normalization factor is added to the formula which is independent from term and document frequencies. We calculated weights of an arbitrary term, w$_{ij}$, using Pivoted Unique Normalization as in equation 2.3.

$$w_{ij} = \frac{\log(dtf) + 1}{sumdtf} \times \frac{U}{1 + 0.0118U} \times \log\left(\frac{N - nf}{nf}\right) \qquad (2.3)$$

where *dtf* is the number of times the term appears in the document, *sumdtf* is the sum of (log(*dtf*)+1)'s for all terms in the same document, *N* is the total number of documents, *nf* is the number of documents that contain the term, *U* is the number of unique terms in the document. The uniqueness means that the measure of document length is based on the unique terms in the document. We used 0.0118 as pivot value. The rank is the product of the weight and the frequency of the term in the query, can be formulated as in equation 2.4.

$$R = \sum_{i=1}^{n} (w_{ij} \times q_i) \qquad (2.4)$$

where *n* is the number of term in the query, $w_{ij}$ is the weight and $q_i$ is the count of term in the query.

## 2.8 Expansion

Expansion techniques are based on the following hypothesis (van Rijsbergen, 1979): “If an index term is good at discriminating relevant from irrelevant documents then any closely associated index term is also likely to be good at this.” When using knowledge structures, expansion terms are determined from pre-fabricated term dependency matrices or lookup tables. Following examples of collection-independent knowledge structures are listed by Efthimiadis (1996):

- Manually constructed, domain-specific thesauri. A thesaurus is a manually crafted or automatically composed list of synonyms or related concepts. It has also been referred to as a "treasury of words" (Foskett, 1997). A thesaurus is domain-specific, if it contains terms from predominantly one particular area, such as medicine or architecture.
- Dictionaries and lexicons, such as Collins dictionary.
- General-purpose thesauri, such as WordNet (Miller, 1990).

Query expansion algorithms based on such references are also known as external techniques as they do not make use of corpus statistics in order to find candidate terms. During query time, queries are expanded simply by looking up related terms in the appropriate structures.

### *2.8.1 Query Expansion*

Under the bag of words model (BOW), if a relevant document does not contain the terms that are in the query, then that document will not be retrieved. The aim of query expansion is to reduce this query/document mismatch by expanding the query using words or phrases with a similar meaning or some other statistical relation to the set of relevant documents.

This procedure may have even greater importance in spoken document retrieval, since the word mismatch problem is heightened by the presence of errors in the automatic transcription of spoken documents. In most collections, the same concept may be referred to using different words. This issue, known as *synonymy*, has an impact SYNONYMY on the recall of most information retrieval systems. For example, you would want a search for *aircraft* to match *plane* (but only for references to an *airplane*, not a woodworking plane), and for a search on thermodynamics to match references to heat in appropriate discussions.

Users often attempt to address this problem themselves by manually refining a query. The methods for tackling this problem split into two major classes: *global methods* and *local methods*. Global methods are techniques for expanding or reformulating query terms independent of the query and results returned from it, so that changes in the query wording will cause the new query to match other semantically similar terms. Global methods include:

- Query expansion/reformulation with a thesaurus or WordNet
- Query expansion via automatic thesaurus generation
- Techniques like spelling correction

Local methods adjust a query relative to the documents that initially appear to match the query. The basic methods here are:

- Relevance feedback
- Pseudo relevance feedback, also known as Blind relevance feedback
- (Global) indirect relevance feedback

Voorhees et al. (1994) expanded queries using WordNet, and found that individual queries that are not well formulated, or do not describe the underlying information need well, can be improved significantly. The results of this work were not good, especially when initial queries are long. In the case of initial short queries, query effectiveness was improved but it was not better than the effectiveness achieved with complete long queries without expansion.

Smeaton et al. (1995) used the concept of specificity of words, and expanded specific terms with the parents and grandparents in the WordNet hierarchy and the abstract terms with the children and grandchildren. Furthermore, every word is expanded with its synonyms. The results in terms of precision were disappointing.

In the work of Mandala et al. (1998) the relations stored in WordNet are combined with similarity measures based on syntactic dependencies and co-occurrence information. This combination improves the effectiveness of retrieval.

The work of Qiu & Frei (1993) used an automatically constructed thesaurus and the results were good but the expansion was tested against small document collections. Other successful Works used thesaurus adapted with relevance information or were tested against collections in specific domains.

### *2.8.2 Document Expansion*

Text retrieval community studied query expansion extensively. However, in literature, document expansion has not been thoroughly researched for information retrieval and especially for ABIR. In document expansion, documents are enriched with related terms. Each document is run as a query and is subsequently expanded with new expansion terms. Search engines generally return documents that contain at least one of the terms in the query. However, Furnas et al. (1987) found that two users, who are asked to describe a certain topic with particular keywords, choose the same keyword with a likelihood of less than 20%.

A technique for updating vector representations is proposed by Ide and Salton (1971). They use relevance feedback, relying on the help of the user. In their work, the query representation is changed to obtain a query vector that is closer to that of the relevant documents.

They also propose a second method, where the document vector space is changed so that relevant documents are closer to the query vector. One of the approaches adds query terms to the vectors of relevant documents. Another approach is to interchange the vector space representation of two documents, one relevant and one non-relevant, with respect to a query. Using these methods, they achieve effectiveness improvements of 10% to 15%.

Exact document expansion, actually adding terms to documents, was first used by Singhal and Pereira (1999) in the context of speech retrieval. Although speech recognition has since improved, at the time of publication of their work, speech recognition was unreliable with an error rate of up to 60%. Singhal and Pereira expand transcribed documents with related terms from a side corpus. This method achieves a relative increase in MAP of 12% over a baseline that was established employing pseudo relevance feedback based on the technique proposed by Rocchio.

Li and Meng (2003) use document expansion for spoken document retrieval, where they expand documents by augmenting them with highly valued tf.idf terms that have been retrieved from a side corpus. Their method is very similar to that of Singhal and Pereira. Li and Meng found a 56% relative improvement in Cantonese monolingual retrieval and 14% relative improvement in Mandarin cross-language retrieval.

In the context of latent semantic indexing, Cristianini et al. (2002) consider "a kind of document expansion" in order to link documents that share related terms. To this end they briefly consider expanding documents by adding all synonyms of terms contained within that document; however, they do not describe any experiments making use of document expansion.

Scholer et al. (2004) augment documents by associating queries obtained from a query log in order to increase retrieval effectiveness. However, they do not reduce the problems of vocabulary mismatch, as they only add queries to documents where all query terms are already part of the document. Instead, they emphasize terms that are central to a document.

The only direct reference to document expansion for document retrieval was made by van Rijsbergen (2000), who pondered whether document expansion could be used in this context. However, no experiments are reported in his paper.

## 2.9 Reranking

Document reranking is a method to reorder the initially retrieved documents with the aim to get better results. We know expansion methods generally results a high recall with low precision in ABIR. Consequently, reranking is fatal for better retrieval.

In literature many reranking approaches has been proposed. The reranking approaches can be roughly classified into several groups based on underlying method used, such as unsupervised document clustering, semi-supervised document categorization, relevance feedback, probabilistic weighting, collaborative filtering or a combination of them. But basically, methods of reranking are grouped under two major approaches. First approach aims to reorder whole result set by using document vectors. In other words, higher ranks are given to relevant documents. These methods are called Reranking methods. Second approach uses pair-wise similarity of retrieved documents instead of term vectors, and aims to model user preferability on document retrieved. These methods are called Learning to Rank methods.

### *2.9.1 Reranking methods*

Reranking methods is done based on the information manifested in the retrieved result set. Relevant documents with low similarity scores are reweighted (by increasing) and reordered.

Carbonell & Goldstein (1998) defines a new criterion for document reranking named maximal marginal relevance (MMR). Goal of this criterion is to eliminate similar document in result set, and present more novelty results to user. Method claims that a relevant document needs to be similar to user query, and need to contain minimal similarity to previously retrieved documents at the same time.

Lingpeng et al. (2005) purposes a document reranking method by applying a weighting scheme on retrieved documents based on MMR. Method uses Chinese documents and six different weighting schemes for retrieved documents. Additionally, method tries to eliminate correlation effect of term while calculating the new weights giving - less weights to terms which can be correlated with a previously weighted term.

Balanski & Danilowicz (2005) uses inter-document similarity information to rerank result set. They use an approximation method to reach ideal document which satisfy use information need. In other words, method tries to find best result set vector which contains documents that are most relevant to user query. Under some assumptions, method uses an iterative algorithm to reduce difference between best document set and result set.

Allan et al. (2001) proposes a clustering method for document reranking. They used an InQuery term weighting scheme proposed by Callan et al. for term weighting. Allan's method uses secant of angle between to document vectors to construct document clusters which is a distance function, 1 / cos $\theta$.

Another clustering approach for document reranking is proposed by Lee et al. (2001). Proposed method defines document similarity by using two similarity scores: classical vector space model score, and cluster analysis score. Method requires document collection to be clustered hierarchically. Cluster analysis performed after initial result set generation for user query. Centroid of the generated result set is calculated and method aims to find closest document cluster for generated result set cluster.

Similarly, an application of Lee's method is performed on image dataset by Park et al. (2005). Image features used in proposed method are color histogram in HSV color space, Gray scale co-occurrence matrixes and edge histograms.

### *2.9.2 Learning to Rank Methods*

The aim these type of methods is to model document which could be preferred than other document by the user. Theoretically, methods model user preference with the help of preference function.

Rigutini et al. (2008) proposes a new learning to rank algorithm to approximate preference function. Proposed method uses a neural network to sort documents in preferable order. Since performance of neural network depends on quality of train set, method adds an incremental training phase to improve performance.

Carvalho et al. (2008) modifies a gradient descend algorithm with a sigmoid loss function to maximize performance of ranking. It is pointed out in the method that using sigmoid loss function reduces effects of outliers in training set.

Metzler & Kanungo (2008) measures performance of some pair-wise similarity methods on automatic document summarization. They use ranking support vector machines (rSVMs), support vector regression (SVG) and gradient boosted decision trees (GBDTs). According to tests GBDT outperforms other method in several datasets.

# CHAPTER THREE
# EXPANSION FOR ANNOTATION BASED IMAGE RETRIEVAL

## 3.1 Introduction

In this chapter, we present our expansion method in detail that is developed for annotation based image retrieval (ABIR) for web images. In proposed system, the aim of expanding both documents and the queries is, to adapt queries to the documents, and documents to queries. We used the same expansion approaches for both documents and queries.

Expanding the poorly defined documents by adding new terms may result in higher ranking performance. Similarly, expanding the queries and widening the search terms increase the recall value by bringing more relevant documents which is not matching literally with the original query. However, there is a risk of constructing more exhaustive documents and queries than original ones with expansion.

In section 3.2, pre-processing phase will be introduced. The details of expansion method of the proposed system will be showed in section 3.3 in which WordNet and related sub-topics (WSD, Similarity functions) will be explained. Finally, an expansion scenario will be illustrated.

## 3.2 Pre-processing

Pre-processing is a kind of data filtering operation. All documents go through this stage before expansion. WikipediaMM task dataset contains 151,519 images and their metadata in XML format. The details of WikipediaMM task and dataset are discussed in Experimentation chapter.

We skipped the useless metadata information in preprocessing step. First, we removed HTML markup tags and special formatting characters. Then, we parsed remaining text and performed the steps below;

- *Case folding:* Case folding, is the process to change all upper case letters into lower case letters or vice versa. This is motivated by the fact that users searching for documents that contain the term “blue flower” are most likely also interested in documents that contain “Blue flower”.

- Removing all punctuations and non-printable characters.

- *Stemming or Lemmatizing using WordNet Lemmatizer:* Stemming is a technique which removes suffixes (Porter, 1980) from terms in order to reduce them to a dictionary form. In inflectional languages, contrary to English, stemming might also remove prefixes or infixes. Stemming typically removes gerunds (“ing”), plurals, and past tenses. In this work, instead of Porter stemming, “WordNet Morphologic Lemmatizer” is used. It works better than Porter stemmer, because stemmed words existence is controlled in WordNet corpora in each step. Table 3.1 presents the different results between Porter Stemmer and WordNet Lemmatizer.

Table 3.1 Different results between Porter Stemmer and WordNet Lemmatizer

| **Porter Stemmer** | **WordNet Stemmer - Lemmatizing** |
|---|---|
| Businesses → busi | Businesses → business |
| Communication → commun | Communication → communication |
| Possible → possibl | Possible → possible |
| Computing → comput | Computing → computing |

- *Stop-words Elimination:* It is a process of removing frequently occurring terms from indexes and queries. The process considers that terms appearing in most documents are not very useful for identifying relevant documents. Although those stopwords have a grammatical function and are important for comprehension of sentences, they are of little use in discriminating some documents from others.

For example, the word “the” occurs in most documents. If “the” was used as part of a query, it would not have a significant impact on the answer set, if any at all. Stop-words include articles, prepositions, and conjunctions; a stoplist may contain 400–500 terms. Stopping process has some advantages: the size of index is reduced by a small percentage and during query evaluation, the inverted lists for stop-words not need be processed, so a considerable time saving is occurred. There are also disadvantages of stopping process: One of them is queries which contain only stop-words, such as “the who”, cannot be serviced with a stopped index. The other disadvantage is that it is difficult to predict exactly which terms will not be of interest to current and future searchers.

- Finally, the documents become available for expansion. Figure 3.1 depicts the UML class diagrams of preprocessing step.

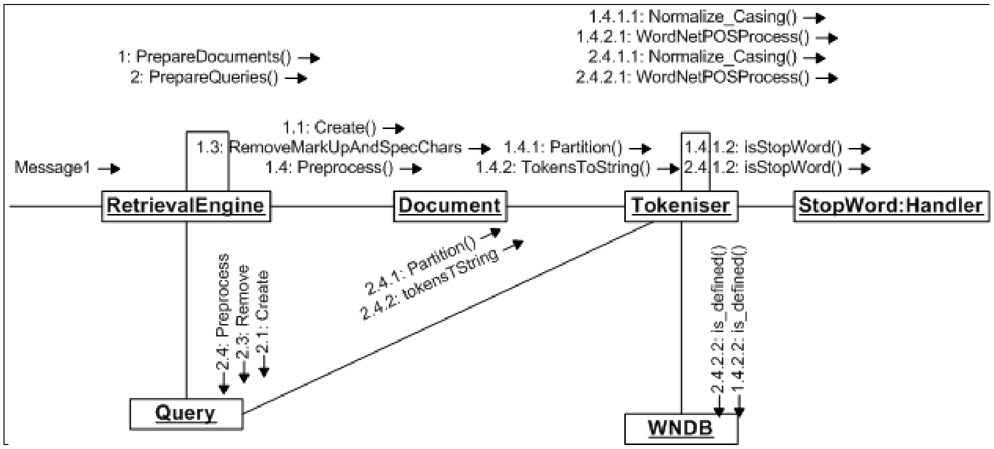


Figure 3.1 UML class diagrams of preparation and preprocessing step.

### 3.3 Expansion Using WordNet

In this thesis, we used WordNet system (Miller, 1990) for both document expansion (DE) and query expansion (QE) steps. WordNet is an on-line lexical reference system developed at Princeton University. WordNet attempts to model the lexical knowledge of a native speaker of English. Word-Net can also be seen as ontology for natural Language terms. It contains around 100,000 terms, organized into taxonomic hierarchies.

Nouns, verbs, adjectives and adverbs are grouped into synonym sets (synsets). The synsets are also organized into senses. The synsets (or concepts) are related to other synsets higher or lower in the hierarchy by different types of relationships. The most common relationships are the *Hyponym/Hypernym* (i.e., is-a relationships), and the *Meronym / Holonym* (i.e., part-of relationships). Although it is commonly argued that language semantics are mostly captured by nouns and noun term-phrases, in this thesis, we considered both noun and adjective representations of terms.

### *3.3.1 WSD (Word Sense Disambiguation) in WordNet*

We used WordNet for Word Sense Disambiguation (WSD) to tune document expansion. Disambiguation is the process of finding out the most appropriate sense of a word that is used in a given sentence. We used an adapted form of a well-known Lesk algorithm (Lesk, 1986) which disambiguates a target word by selecting the sense whose dictionary gloss shares the largest number of words with the glosses of neighboring words.

The original Lesk algorithm uses dictionary definitions (gloss) to disambiguate a polysemous word in a sentence context. The major objective of his idea is to count the number of words that are shared between two glosses. The more overlapping the words, the more related the senses are. The algorithm begins a new for each word and does not utilize the senses it previously assigned. This greedy method does not always work effectively. The major idea behind such methods is to reduce the search space by applying several heuristic techniques. The Beam searcher limits its attention to only *k* most promising candidates at each stage of the search process, where *k* is a predefined number. The adapted Lesk algorithm (Banerjee & Pederson, 2003) is described in the following steps:

1. Select a context: optimizes computational time so if *N* is long, *K* context will be defined around the target word (or k-nearest neighbor) as the

sequence of words starting *K* words to the left of the target word and ending *K* words to the right. This will reduce the computational space that decreases the processing time. For example: If *K* is four, there will be two words to the left of the target word and two words to the right.

2. For each word in the selected context, all the possible senses are listed whit their POS (part of speech) noun and verb.

3. For each sense of a word (WordSense), following relations (example of pine and cone) are listed:
    - Its own gloss/definition that includes example texts that WordNet provides to the glosses.
    - The gloss of the synsets that are connected to it through the hypernym relations. If there is more than one hypernym for a word sense, then the glosses for each hypernym are concatenated into a single gloss string (*).
    - The gloss of the synsets that are connected to it through the hyponym relations (*).
    - The gloss of the synsets that are connected to it through the meronym relations (*).
    - The gloss of the synsets that are connected to it through the troponym relations (*).

    (*) All of them are applied with the same rule.

4. Combine all possible gloss pairs that are archived in the previous steps and compute the relatedness by searching for overlap. The overall score is the sum of the scores for each relation pair. To score the overlap a new scoring mechanism is used that differentiates between N-single words and N-consecutive word overlaps and effectively treats each gloss as a bag of words. It is based on ZipF's Law, which says that the length of words is inversely proportional to their usage. The shortest words are those which are used more often, the longest ones are used less often. Measuring

overlaps between two strings is reduced to solve the problem of finding the longest common sub-string with maximal consecutives. Each overlap which contains *N* consecutive words, contributes *N2* to the score of the gloss sense combination.

5. Once each combination has been scored, the sense that has the highest score is picked up to be the most appropriate sense for the target word in the selected context space. Hopefully the output not only gives us the most appropriate sense but also the associated part of speech for a word. If you intend to work with this topic, you should refer to the measurements of Hirst-St.Onge which is based on finding the lexical chains between the synsets.

To disambiguate a word, the gloss of each of its senses is compared to the glosses of every other word in a phrase. A word is assigned to the sense whose gloss shares the largest number of words in common with the glosses of the other words. For example: In performing disambiguation for the "pine cone" phrasal, according to the Oxford Advanced Learner’s Dictionary, the word "pine" has two senses:

- sense 1: kind of evergreen tree with needle–shaped leaves,
- sense 2: waste away through sorrow or illness.

The word "cone" has three senses:

- sense 1: solid body which narrows to a point,
- sense 2: something of this shape whether solid or hollow,
- sense 3: fruit of a certain evergreen tree.

By comparing each of the two gloss senses of the word "pine" with each of the three senses of the word "cone", it is found that the words "evergreen tree" occurs in one sense in each of the two words. So these two senses are then declared to be the most appropriate senses when the words "pine" and "cone" are used together. Figure

3.2 presents the UML class diagrams of WordNet library and packages in proposed system.

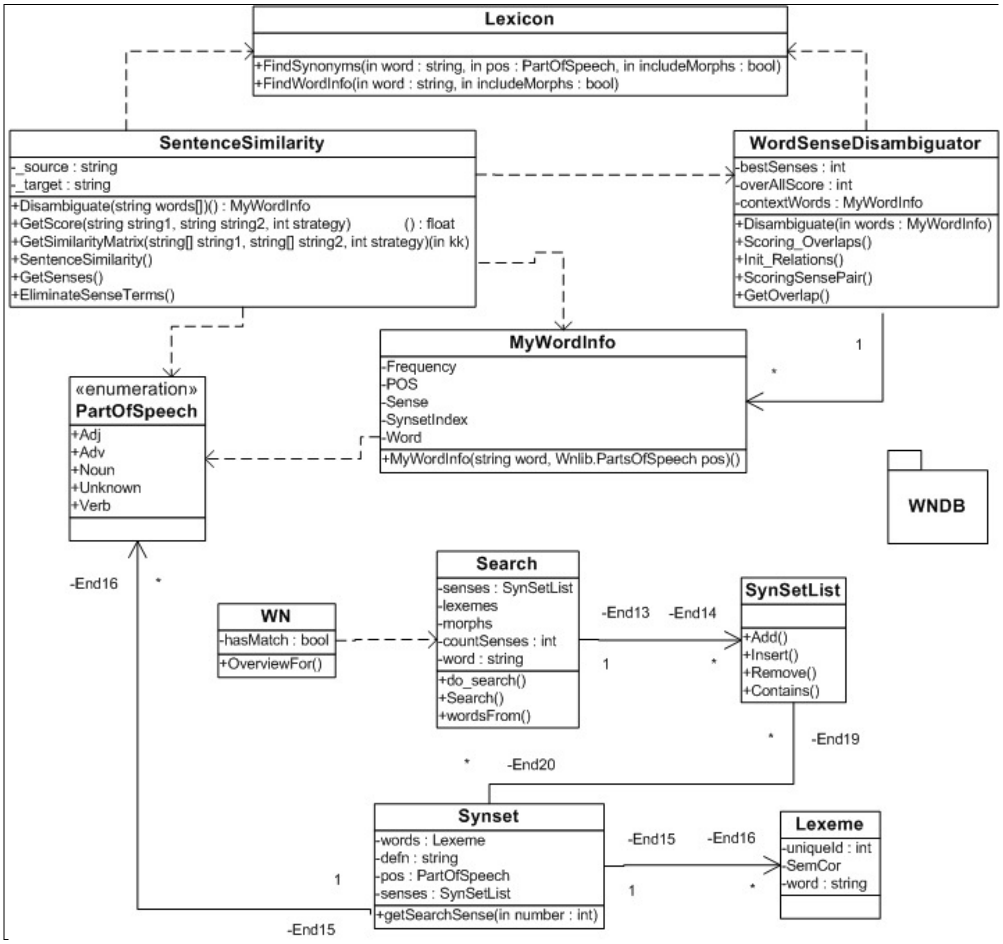


Figure 3.2 UML class diagrams of WordNet library and packages.

### *3.3.2 Semantic Similarity in WordNet*

Several methods for determining semantic similarity between terms have been proposed in the literature. Similarity measures apply only for nouns and verbs in WordNet (taxonomic properties for adverbs and adjectives do not exist). Semantic similarity methods are classified into four main categories:

a. *Edge Counting Methods:* Measure the similarity between two terms (concepts) as a function of the length of the path linking the terms and on the position of the terms in the taxonomy.

b. *Information Content Methods:* Measure the difference in information content of the two terms as a function of their probability of occurrence in a corpus.

c. *Feature based Methods:* Measure the similarity between two terms as a function of their properties (e.g., their definitions or ”glosses” in WordNet) or based on their relationships to other similar terms in the taxonomy

d. *Hybrid methods:* combine the above ideas.

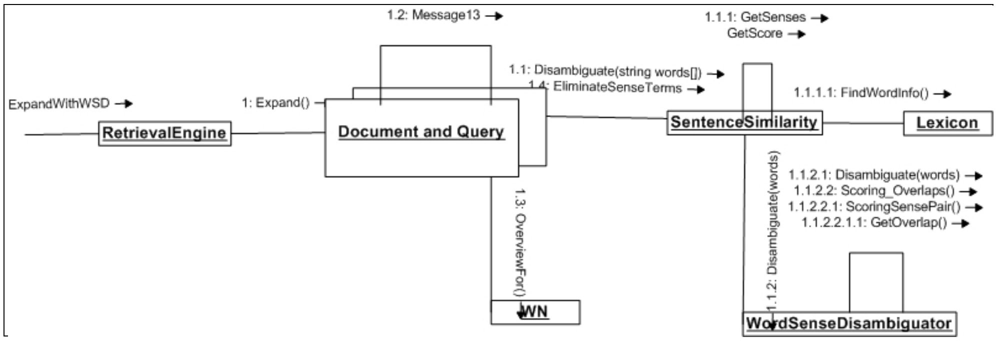

1.2: Message13 →
1.1.1: GetSenses →
GetScore →
1.1: Disambiguate(string words[]) →
1.4: EliminateSenseTerms →
1.1.1.1: FindWordInfo() →
ExpandWithWSD →
1: Expand()
RetrievalEngine
Document and Query
SentenceSimilarity
Lexicon
1.1.2.1: Disambiguate(words) →
1.1.2.2: Scoring_Overlaps() →
1.1.2.2.1: ScoringSensePair() →
1.1.2.2.1.1: GetOverlap() →
1.3: OverviewFor()
1.1.2: Disambiguate(words)
WN
WordSenseDisambiguator

Figure 3.3 UML class diagrams and packages of WordNet WSD and similarity process.

To measure the semantic similarity between two synsets Hyponym/hypernym (or is-a relations) is used. A simple way to measure the semantic similarity between two synsets is to treat taxonomy as an undirected graph and measure the distance between them in WordNet. The shorter the path from one node to another, the more similar they are. Note that the path length is measured in nodes/vertices rather than in links/edges. The length of the path between two members of the same synsets is 1 (synonym relations). Figure 3.4 shows an example of the hyponym taxonomy in WordNet used for path length similarity measurement:

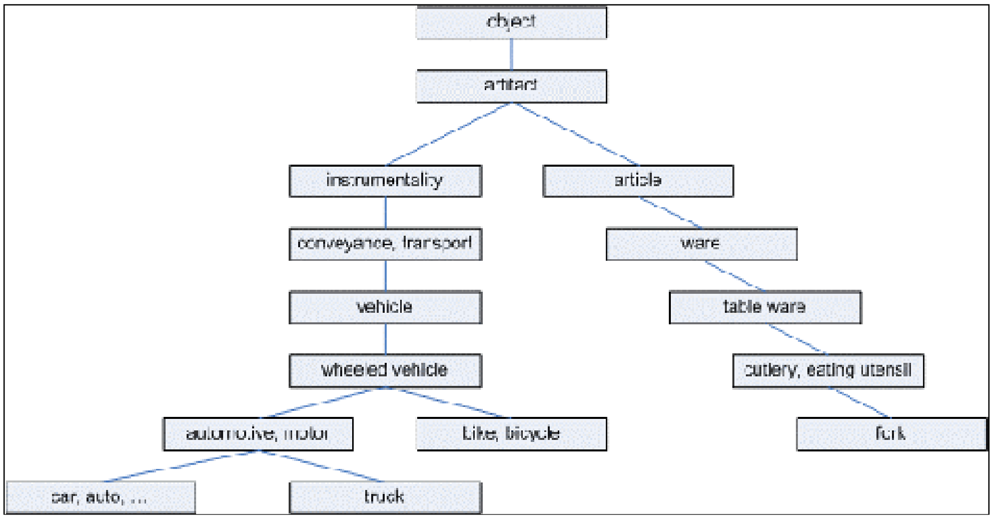


Figure 3.4 Sample hyponym taxonomy in WordNet.

It is observed that the length between car and auto is 1, car and truck is 3, car and bicycle is 4, car and fork is 12. A shared parent of two synsets is known as a sub-sumer. The least common sub-sumer (LCS) of two synsets is the sumer that does not have any children that are also the sub-sumer of two synsets. In other words, the LCS of two synsets is the most specific sub-sumer of the two synsets. Back to the above example, the LCS of {car, auto..} and {truck..} is {automotive, motor vehicle}, since the {automotive, motor vehicle} is more specific than the common sub-sumer {wheeled vehicle}.

Measuring similarity (MS1 – Shortest Path Length): There are many proposals for measuring semantic similarity between two synsets: Leacock & Chodorow, P.Resnik.

$$Sim(s,t) = 1/distance(s,t) \qquad (3.1)$$

where distance is the shortest path length from *s* to *t* by using node counting.

Measuring similarity (MS2 – Wu & Palmer Method): This formula is proposed by Wu & Palmer, the measure considers both path length and depth of the least common sub-summer as in eq. 3.2.

$$Sim(s,t) = 2 * depth(LCS)/[depth(s) + depth(t)] \qquad (3.2)$$

where $s$ and $t$ denote the source and target words being compared. Depth($s$) is the shortest distance from root node to a node $S$ on the taxonomy where the synset of $S$ lies. LCS denotes the least common sub-sumer of $s$ and $t$.

### *3.3.3 Expansion Scenario*

Figure 3.5 illustrates the expansion of query numbered 1 in WikipediaMM task. The query *"blue flowers"* is firstly preprocessed, terms *"blue"* and *"flower"* are generated. Each term's senses are fetched from WordNet. In our query, *"blue"* has 7 senses and *"flower"* has 3 senses. Since numerous senses exists in different domains for terms, expanding the term with all of these senses results too noisy exhaustive documents/queries. We prevent such noisy expansions by selecting the most appropriate sense with Lesk's WSD algorithm.

In our example, first senses of terms are selected. In WordNet, a sense consists of two parts; synonym words and sense definition. We used both of them for expansion in our work. We again preprocess the whole parts of selected sense to reduce noise level. Then, we check the expanded terms' existence in dataset dictionary and if it not exists we eliminate them. At the end of this rule, *"flower"* has new expanded terms; *plant*, *cultivated*, *blossom* and *bloom*. For each one, we calculate a similarity score between their base terms (flower).

As discussed before, in literature, different methods have been proposed to measure the semantic similarity between terms (Wu & Palmer, 1994; Richardson, Smeaton, & Murphy, 1994; Li, Bandar, & McLean, 2003; Resnik, 1999; Tversky, 1977). In this thesis, we used Wu and Palmer's edge counting method (Wu & Palmer, 1994).

Finally, we add terms above a specific threshold value to the final query or document. Threshold values for noun and adjective terms are 0.9 and 0.7, respectively. In our example, query *“blue flower”* is finally expanded as *“blue flower blueness sky bloom blossom”*.

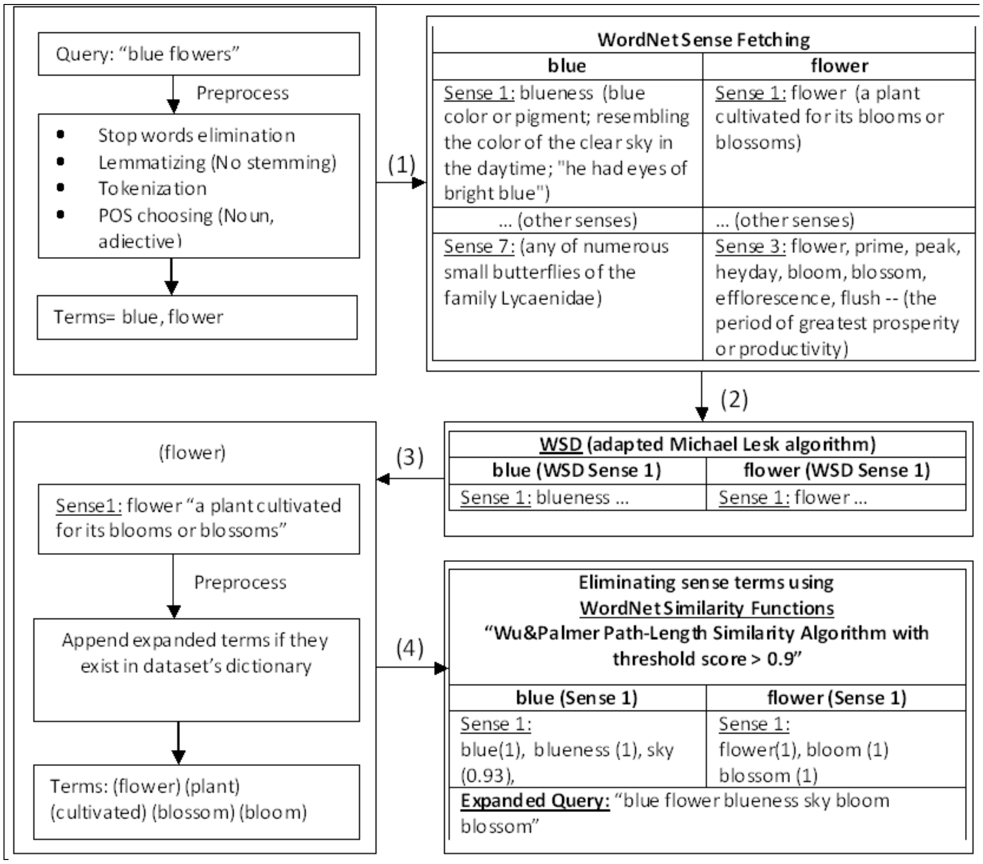

Query: “blue flowers”
Preprocess
Stop words elimination
Lemmatizing (No stemming)
Tokenization
POS choosing (Noun, adjective)
Terms= blue, flower
(1)
WordNet Sense Fetching
blue
flower
Sense 1: blueness (blue color or pigment; resembling the color of the clear sky in the daytime; "he had eyes of bright blue")
Sense 1: flower (a plant cultivated for its blooms or blossoms)
... (other senses)
... (other senses)
Sense 7: (any of numerous small butterflies of the family Lycaenidae)
Sense 3: flower, prime, peak, heyday, bloom, blossom, efflorescence, flush -- (the period of greatest prosperity or productivity)
(2)
WSD (adapted Michael Lesk algorithm)
blue (WSD Sense 1)
flower (WSD Sense 1)
Sense 1: blueness ...
Sense 1: flower ...
(3)
(flower)
Sense1: flower “a plant cultivated for its blooms or blossoms”
Preprocess
Append expanded terms if they exist in dataset’s dictionary
Terms: (flower) (plant) (cultivated) (blossom) (bloom)
(4)
Eliminating sense terms using WordNet Similarity Functions
“Wu&Palmer Path-Length Similarity Algorithm with threshold score > 0.9”
blue (Sense 1)
flower (Sense 1)
Sense 1: blue(1), blueness (1), sky (0.93),
Sense 1: flower(1), bloom (1) blossom (1)
Expanded Query: “blue flower blueness sky bloom blossom”

Figure 3.5 Sample query expansion.

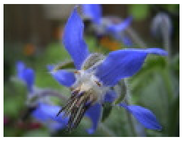
Term phrase selection (TPS) is one of the major parts of expansion phase. During expansion, we checked every successive word pairs for existence in WordNet as noun-phrase. If it exists, we expanded document/query by appending term phrase is appended to the dictionary. For example, if a document contains “hunting dog”, these two successive tokens are searched in WordNet. If this phrase exists, the document is expanded with the term “hunting dog”. Finally the term phrase is added to the term phrase dictionary.

For Wikipedia collection, the numbers of new term-phrases added was 6,808. Some term-phrase examples are railway station, great hall, Forbidden City, colonel blimp, web site, limited edition, riot gun, web browser, bank note, red bay, saint thomas. 87 term-phrase sample can be found at the appendix section A.3.

Table 3.2 depicts the same query number 1 and its two relevant documents with ID of 1027698 and 163477 by showing their original and expanded forms. Relevant documents are about some kind of flowers that are uploaded to Wikipedia pages. The query is *blue flowers*. Both *borage* and *lavender* are somehow related with *blue flowers* although their documents don't include these terms. In such cases, without any expansion technique, retrieval performance will not be satisfactory.

The example also shows that expanding query only is not adequate, where only the terms of *blueness*, *sky*, *bloom* and *blossom* are added to query. However; we must also expand the documents to match. After document expansion, the terms *blue* and *flower* are added to both documents. In addition to this, *bloom* and *blossom* terms are also appended to document numbered 163477. As a result, the expansion step adds new common terms to both documents and query by using WordNet, WSD. Then, whole VSM is rebuilt based on new dictionary.

Table 3.2 Expanded document and query samples

| **Image/ Query ID** | **Image** | **Original Document / Query** | **Expanded Document / Query** |
|---|---|---|---|
| Doc #:1027698 | | sea lavender limonium | sea lavender limonium sealavender statice various plant genus limonium temperate salt marsh spike whit mauve **flower** various old world aromatic shrub subshrub mauve **blue_**cultivated division ocean body salt water enclosed land |
| Doc #:163477 |  | borage flower garden made apr | borage flower garden made apr made plant cultivated **bloom blossom** tailwort hairy **blue flowered** european annual herb herbal medicine raw salad greens cooked spinach april month preceding plot ground plant cultivated |
| Query #:1 | N/A | blue flower | blue flower blueness sky **bloom blossom** |

After describing proposed expansion methods for ABIR, we can give some preliminary definitions and formulas. We perform an initial retrieval, called *base result* using well known vector space model (VSM) and pivoted unique normalization. The formula to calculate the base similarity score is presented as in equation 3.3:

$$r(j) = \sum_{i=1}^{n} (w_{ij} \times q_i) \qquad (3.3)$$

where, $r(j)$ is the similarity score of $j^{th}$ document, *n* is length of the vocabulary and, $w_{ij}$ and $q_i$ represents the weight of $j^{th}$ document and $i^{\text{th}}$ query, respectively. Let us assume that $\acute{r}(j)$ represents the similarity score of *expanded documents*. To calculate overall similarity score, we use both *expanded* and *original* similarity scores by taking the averages of them with some coefficients, formulated as in equation 3.4.

$$R_0(j) = \frac{(r(j) \times \mu) + (\acute{r}'(j) \times \partial)}{2} \qquad (3.4)$$

where, $R_0(j)$ show the initial similarity score of $j^{\text{th}}$ document, $\mu$ and $\partial$ are coefficients to adjust results for different datasets and queries. In this study, we empirically set $\mu$ and $\partial$ values to 1 and 0.9, respectively.

# CHAPTER FOUR

# TWO-LEVEL RERANKING FOR ANNOTATION BASED IMAGE RETRIEVAL

## 4.1 Introduction

To improve precision at the top ranks of results returned for a query, researchers suggested to automatically rerank the documents in an initially retrieved list as described in Chapter 2.

In this thesis, we propose a new reranking method which includes two-level: The first level forms a narrowing-down phase of search space while second level includes a cover coefficient based reranking. Narrowing-down level will be described in section 4.2. In section 2.4, preprocessing step will be explained. In section 4.3, cover coefficient based reranking step and C3M algorithm will be described and demonstrated in detail respectively.

## 4.2 First Level: Narrowing-down

The first level of our reranking approach forms a narrowing-down phase and includes re-indexing. Result sets of each query and corresponding base similarity scores, $R_0(j)$, are inputs for reranking operation.

Table 4.1 First level reranking initial result sets (Shrink down)

| Result Set ID | Query ID | # of Images in Result Set |
|---|---|---|
| Dataset #:1 | Query #:1 | 480 |
| Dataset #:2 | Query #:2 | 340 |
| … | … | … |
| Dataset #:m | Query #:m | 500 |

In this level we first select relevant documents using initial similarity scores, $R_0$. In other words, we filter out non-relevant documents based on initial similarity scores. This operation drastically reduces both the number of documents and the number of terms.

Table 4.2 Query retrieval scores for new result sets in first level reranking step

| **Result Set ID** | **Query ID** | **Image ID** | **Base Ranking Score, $R_0(j)$** | **New Ranking Score, $r_1(j)$** |
|---|---|---|---|---|
| Dataset #:1 | Query #:1 | 1027698 | 3.122 | 1.248 |
| Dataset #:1 | Query #:1 | 163477 | 4.1664 | 0.986 |
| … | … | … | … | … |
| Dataset #:2 | Query #:2 | … | … | … |
| … | … | … | … | … |
| Dataset #:m | Query #:m | … | ... | … |

Then we constructed a new VSM using this small document sets. In short, this level shrinks down the initial VSM data into more manageable size so that we perform more complex cover coefficient based reranking algorithm upon. Table 4.1 presents the number of images/documents after performing base retrieval for each query.

For example for query #1 (blue flower), we have 480 results after base retrieval which is the new compact VSM dataset. This process is done for all queries and new VSM datasets are constructed as the number of queries. Base retrieval similarity scores are also kept to calculate new ranking score. Here after, we calculated first level similarity scores, $R_1(j)$, formulated as in equation 4.1.

$$R_1(j) = (R_0(j) \times \alpha) + r_1(j) + \beta \qquad (4.1)$$

where $r_1(j)$ is the new similarity score of $j^{th}$ document in new small VSM. The value of $\alpha$ is the weight factor and empirically set to 0.8. Additionally, $\beta$ is set to 4 if $j^{\text{th}}$document contains the original query terms in exact order, zero otherwise.

Table 4.3 presents some final similarity scores after the first level reranking step is performed. First level reranking similarity score $(R_1(j))$ for the Image-1027698 is 3.487. This is a better similarity score than base ranking score $(R_0(j))$ 3.122. So this image will be reranked (reordered) in the result set. As a result, precision measurement will be improved since it is a relevant document for the query #1.

Table 4.3 Final similarity scores sample table after first level reranking step

| Dataset ID | Query ID | Image ID | Base Ranking Score (Initial), $R_0(j)$ | New Ranking Score, $r_1(j)$ | First Level: Reranking Score, $R_1(j)$ |
|---|---|---|---|---|---|
| Dataset #:1 | Query #:1 | 1027698 | 3.122 | 1.248 | 3.487 |
| Dataset #:1 | Query #:1 | 163477 | 4.1664 | 0.986 | 8.297 |
| … | … | … | … | … | … |
| Dataset #:2 | Query #:2 | … | … | … | … |
| … | … | … | … | … | … |
| Dataset #:m | Query #:m | … | ... | … | … |

Figure 4.1 presents the general view of first level narrowing-down reranking approach of proposed system.

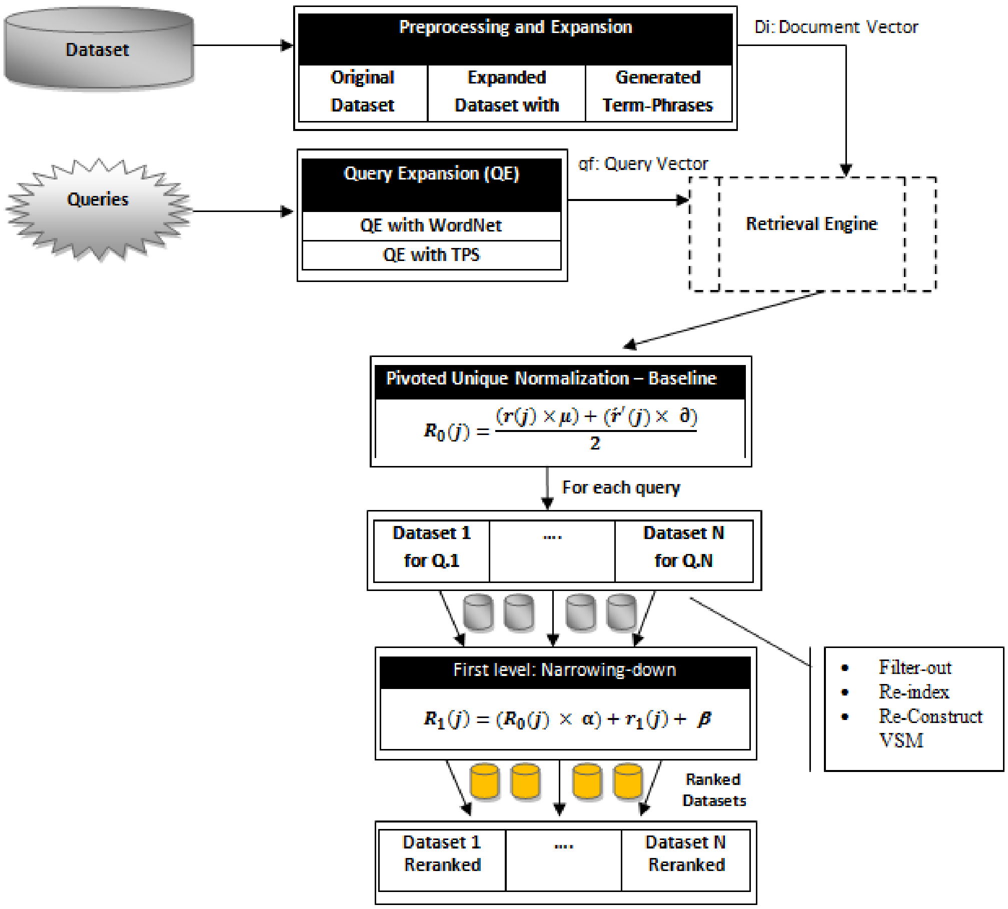


Figure 4.1 First level reranking: narrowing-down view

### 4.3 Second Level: Cover Coefficient-based

In the second level, we propose Cover Coefficient based reranking method. Concept of Cover Coefficient is originally proposed by (Can & Ozkarahan, 1990) for text clustering. CC concept provides a means of estimating the number of clusters within a document database and relates indexing and clustering analytically. The CC concept is used also to identify the cluster seeds and to form clusters with these seeds. The retrieval experiments show that the information retrieval effectiveness of the algorithm is compatible with a very demanding complete linkage clustering method that is known to have good retrieval performance.

Cover Coefficient-based Clustering Methodology ($C^3M$) employs document clusters as cluster seeds and member documents. Cluster seeds are selected by employing the seed power concept and the documents with the highest seed power are selected as the seed documents.

In their paper Can, F. and Ozkarahan E.A., they showed that the complexity of $C^3M$ is better than most other clustering algorithms, whose complexities range from $O(m^2)$ to $O(m^3)$. Also their experiments show that $C^3M$ is time efficient and suitable for very large databases. Its low complexity is experimentally validated. $C^3M$ has all the desirable properties of a good clustering algorithm.

$C^3M$ algorithm is a partitioning type clustering in which clusters cannot have common documents. A generally accepted strategy to generate a partition is to choose a set of documents as the seeds and to assign the ordinary (non-seed) documents to the clusters initiated by seed documents to form clusters. This is the strategy used by $C^3M$. Cover coefficient, CC, is the base concept of $C^3M$ clustering.

The CC concept serves to;

i. identify relationships among documents of a database by use of the CC matrix,

ii. determine the number of clusters that will result in a document database;

iii. select cluster seeds using a new concept, cluster seed power;

iv. form clusters with respect to C$^3$M, using concepts (i)-(iii);

v. correlate the relationships between clustering and indexing.

C$^3$M is a seed-based partitioning type clustering scheme. Basically, it consists of two different steps that are cluster seed selection and the cluster construction. *D* matrix is the input for C$^3$M, which represents documents and their terms. It is assumed that each document contains *n* terms and database consists of *m* documents.

### *4.3.1 C Matrix*

The need is to construct *C* matrix, in order to employ cluster seeds for C$^3$M. *C*, is a document-by-document matrix whose entries $c_{ij}$ $(1 < i, j < m)$ indicate the probability of selecting any term of $d_i$ from $d_j$. In other words, the *C* matrix indicates the relationship between documents based on a two-stage probability experiment.

The experiment randomly selects terms from documents in two stages. The first stage randomly chooses a term $t_k$ of document $d_i$; then the second stage chooses the selected term $t_k$ from document $d_j$. For the calculation of *C* matrix, $c_{ij}$, one must first select an arbitrary term of $d_i$, say, $t_k$, and use this term to try to select document $d_j$ from this term, that is, to check if $d_j$ contains $t_k$. Each row of the *C* matrix summarizes the results of this two-stage experiment.

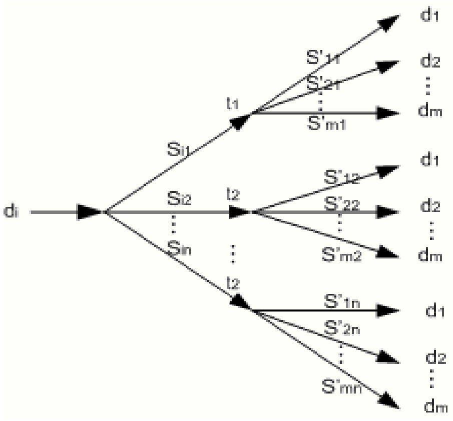


Figure 4.2 Hierarchical representation of two stage probability model for $d_i$ of D Matrix.

Let $s_{ik}$ indicate the event of selecting $t_k$ from $d_i$ at the first stage, and let $s'_{jk}$ indicate the event of selecting $d_j$, from $t_k$ at the second stage. In this experiment, the probability of the simple event "$s_{ik}$ and $s'_{jk}$" that is, $P(s_{ik}, s'_{jk})$ can be represented as $P(s_{ik})\times\ P(s'_{jk})$ . To simplify the notation, we use $s_{ik}$ and $s'_{jk}$ respectively as in equation 4.2.a and 4.2.b, for $P(s_{ik})$ and $P(s'_{jk})$.

$$s_{ik} = \frac{d_{ik}}{\sum_{h=1}^{n}(d_{ih})}, \text{ where } 1 \le i,\ \ j \le m, 1 \le k \le n \qquad (4.2.a)$$

$$s'_{jk} = \frac{d_{jk}}{\sum_{h=1}^{m}(d_{hk})}\ , \text{ where } 1 \le i,\ \ j \le m, 1 \le k \le n \qquad (4.2.b)$$

By considering document $d_i$, $D$ matrix can be represented with respect to the two-stage probability model. Each element of $C$ matrix, $c_{ij}$, (the probability of selecting a term of $d_i$ from $d_j$) can be founded by summing the probabilities of individual path from $d_i$ to $d_j$. $c_{ij}$ can be formulated as in equation 4.3.

$$c_{ij} = \sum_{i=1}^{n}\left(s_{ik}\ \times\ s'_{jk}\right) \qquad (eq{:}4.3)$$

To decrease the complexity of calculating$c_{ij}$, this can be written as in eq. 4.4;

$$(c_{ij} = \alpha_i \ \sum_{k=1}^{n}(d_{ik} \times \beta_k \times d_{jk}), \text{ where } 1 \leq i,\ j \leq m) \qquad (4.4)$$

where $\alpha_i$ and $\beta_k$ are reciprocals of the $i$ th row sum and $k^{th}$ column sum, respectively, as shown in eq. 4.5 and 4.6.

$$\alpha_i = \frac{1}{\sum_{j=1}^{n}(d_{ij})}\ \text{, where, } 1 \leq i \leq m \qquad (4.5)$$

$$\beta_k = \frac{1}{\sum_{j=1}^{m}(d_{jk})}\ \text{, where, } 1 \leq k \leq n \qquad (4.6)$$

The following properties hold for the C matrix:

i. For $i \neq j$ , $0 \leq c_{ij} \leq c_{ii}$ and $c_{ii} > 0$

ii. $c_{i1} + c_{i2} + c_{i3} + ... + c_{im} = 1$

iii. If none of the terms of $d_i$ is used by the other documents, then $c_{ii}=1$ otherwise, $c_{ii}<1$.

iv. If $c_{ij} = 0$, then $c_{ji} = 0$, and similarly, if $c_{ij} > 0$, then $c_{ji} > 0$; but in general, $c_{ij} \neq c_{ji}$.

v. $c_{ii} = c_{jj}, = c_{ij} = c_{ji}$ iff $d_i$ and $d_j$ are identical.

From these properties of the C matrix and from the CC relationships between two document vectors, $c_{ij}$, can be seen as in eq. 4.7.

$$c_{ij} = \begin{cases} \text{extent to which } d_i \text{ covered by } \mathrm{d_j} \text{ for } i \neq j \\ \text{(coupling of } d_i \text{ with } d_j\text{)} \\ \text{extent to which } d_i \text{ covered by itself for } i = j \\ \text{(decoupling of } d_i \text{ from the rest of documents)} \end{cases} \qquad (4.7)$$

To obtain a better understanding of the meaning of the C matrix, consider two document vectors $d_i$ and $d_j$. For these document vectors, four possible relationships can be defined in terms of C matrix entries:

- *Identical documents:* Coupling and decoupling of any two such documents are equivalent. Furthermore, the extent to which these two documents are covered by documents is also identical.

- *Overlapping documents:* Each document will cover itself more than any other ($c_{ii} > c_{ij}$ , $c_{jj} > c_{ji}$). However, this does not provide enough information to compare $c_{ii}$with $c_{jj}$ and $c_{ij}$with $c_{ji}$.

- *A document is a subset of another document:* Let $d_i$ be a subset of $d_j$. Since $d_i$ is a subset of $d_j$ the extent to which $d_i$ is covered by itself will be identical to the extent to which di is covered by $d_j$.

- *Disjoint documents:* Since $d_i$ and $d_j$ do not have any common terms, then they will not cover each other.

As can be seen from the foregoing discussions, in a D matrix, if $d_i$ is relatively more distinct, then $c_{ii}$ will take higher values. Because of this, $c_{ii}$ is called the decoupling coefficient, $\delta_i$, of $d_i$. The sum of the off-diagonal entries of the ith row indicates the extent of coupling of $d_i$ with the other documents of the database and is referred to as the coupling coefficient, $\psi_i$, of $d_i$. From the properties of C matrix,

$\delta_i = c_{ii}$ : decoupling coefficient of $d_i$

$\psi_i = 1 - \delta_i$ : coupling coefficient of $d_i$

### *4.3.2 Reranking Method*

Figure 4.3 shows the general view of the second level cover coefficient based reranking method and final ranking calculation formula of the proposed system.

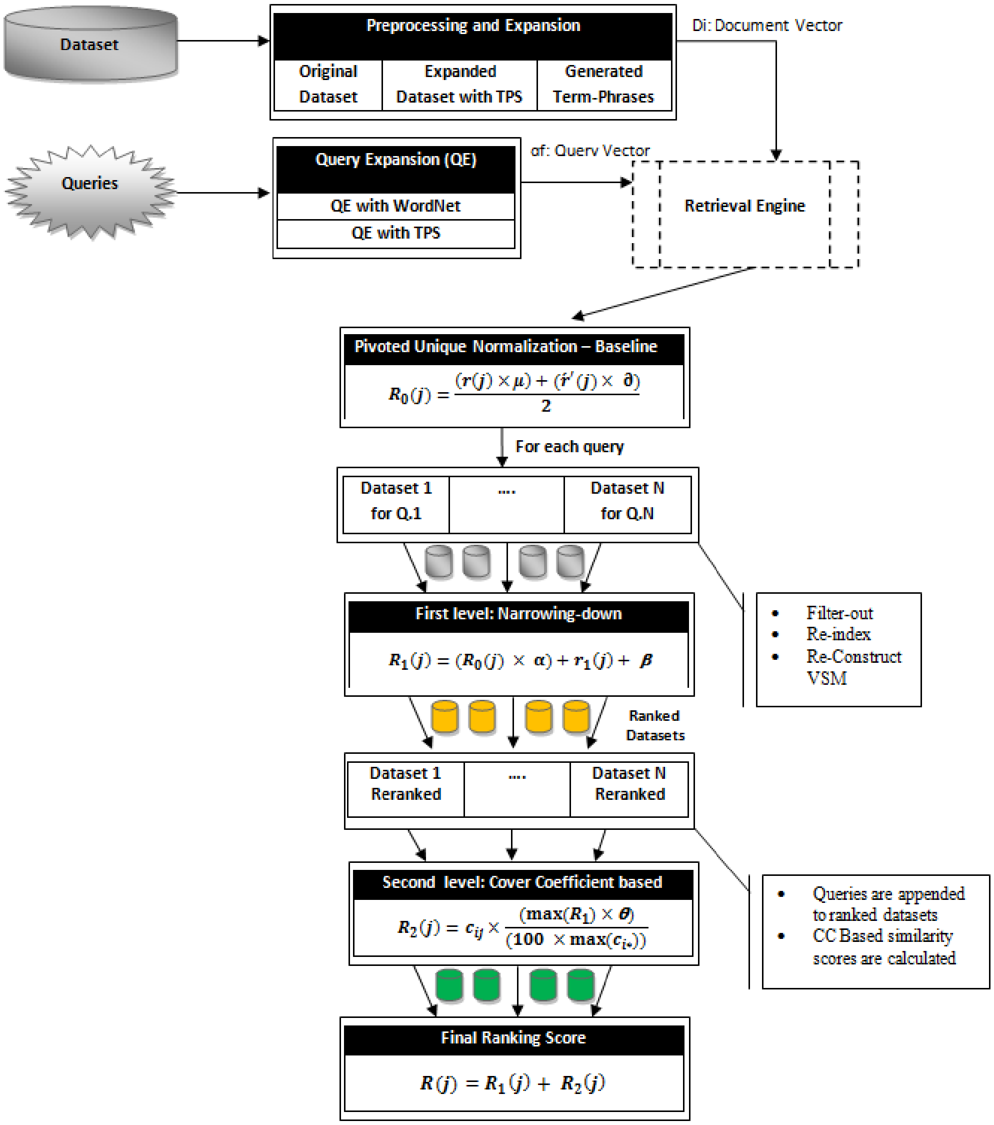

Dataset
Preprocessing and Expansion
Original Dataset
Expanded Dataset with TPS
Generated Term-Phrases
Di: Document Vector
Queries
Query Expansion (QE)
QE with WordNet
QE with TPS
qf: Query Vector
Retrieval Engine
Pivoted Unique Normalization – Baseline
$R_0(j) = \frac{(r(j) \times \mu) + (\acute{r}'(j) \times \partial)}{2}$
For each query
Dataset 1 for Q.1
....
Dataset N for Q.N
First level: Narrowing-down
$R_1(j) = (R_0(j) \times \alpha) + r_1(j) + \beta$
Filter-out
Re-index
Re-Construct VSM
Ranked Datasets
Dataset 1 Reranked
....
Dataset N Reranked
Second level: Cover Coefficient based
$R_2(j) = c_{ij} \times \frac{(\max(R_1) \times \theta)}{(100 \times \max(c_{i*}))}$
Queries are appended to ranked datasets
CC Based similarity scores are calculated
Final Ranking Score
$R(j) = R_1(j) + R_2(j)$

Figure 4.3 Second level reranking and final ranking score calculation

In order to perform reranking, we first appended the query into new VSM as a document, and then calculated $C$ matrix as described above. The $C$ matrix entries, $c_{ij}$, show how $i^{th}$ document is covered by $j^{th}$ document. We considered $i^{th}$ row of $C$

matrix, which includes how query is covered by other documents. We calculated final similarity score using both $R_1(j)$ and $c_{ij}$ as in equation 4.8.

$$R_2(j) = c_{ij} \times \frac{(\max(R_1) \times \theta)}{(100 \times \max(c_{i*}))} \qquad (4.8)$$

where, $max(R^b)$ is the maximum first level similarity score for the query result set, $\theta$ is an empirical coefficient that specifies the percentage of similarity score effect, $\max(c_{i*})$ is the maximum query-by-document similarity score for the $i^{\text{th}}$query. Finally, CC based similarity score equation is as follows:

$$R(j) = R_1(j) + R_2(j) \qquad (4.9)$$

where, $R_1(j)$ is the first level, $R_2(j)$ is the second level and $R(j)$ is the final similarity scores to be used to calculate ranking scores.

# CHAPTER FIVE
# EXPERIMENTATIONS AND EVALUATIONS

## 5.1 Introduction

In the previous chapters we have described the details of our expansion and reranking method for ABIR from web. There is a need to make a performance evaluation for the method in order to show the effectiveness and to make benchmarking. Performance evaluation is not a trivial task because of it's visually inspection has large number of images which is not being practical. An application has been developed to make performance evaluation for the proposed method's experiments.

In this chapter we will describe our System View in section 5.2. Then we will express the ImageCLEF WikipediaMM subtask we participated, and web based data set we have used in our retrieval experiments in section 5.3 and 5.4. In section 5.5, 5.6 and 5.7, we will express evaluation measures, database structure and development environment of proposed system. Finally, we will show the experimental results of annotation based image retrieval techniques that are based on expansion and reranking to evaluate the strengths of our methods in section 5.8.

## 5.2 System View

Initially, system performs basic preprocessing such as punctuation deletion, stop-word elimination and lemmatizing etc. Then, documents are expanded using WordNet and term phrases are selected. During this phase, we take into consideration both the original and the expanded forms of dataset to calculate similarity score and converted document vectors before base retrieval. Figure 5.1 presents the general UML class diagrams of proposed system.

Besides, we also expand queries using TPS and/or WordNet for experimental purposes. During the base retrieval, system uses a combination of similarity scores on expanded and original datasets, then stores each query result set temporarily with the form of Pivoted Unique Normalization (Garcia, 2006), that is a modified version of the classical cosine normalization (Salton, Wong, & Yang, 1975) based on term weighting aspect of modern text retrieval systems (Buckley, 1993; Manning, Raghavan, & Schütze, 2009).

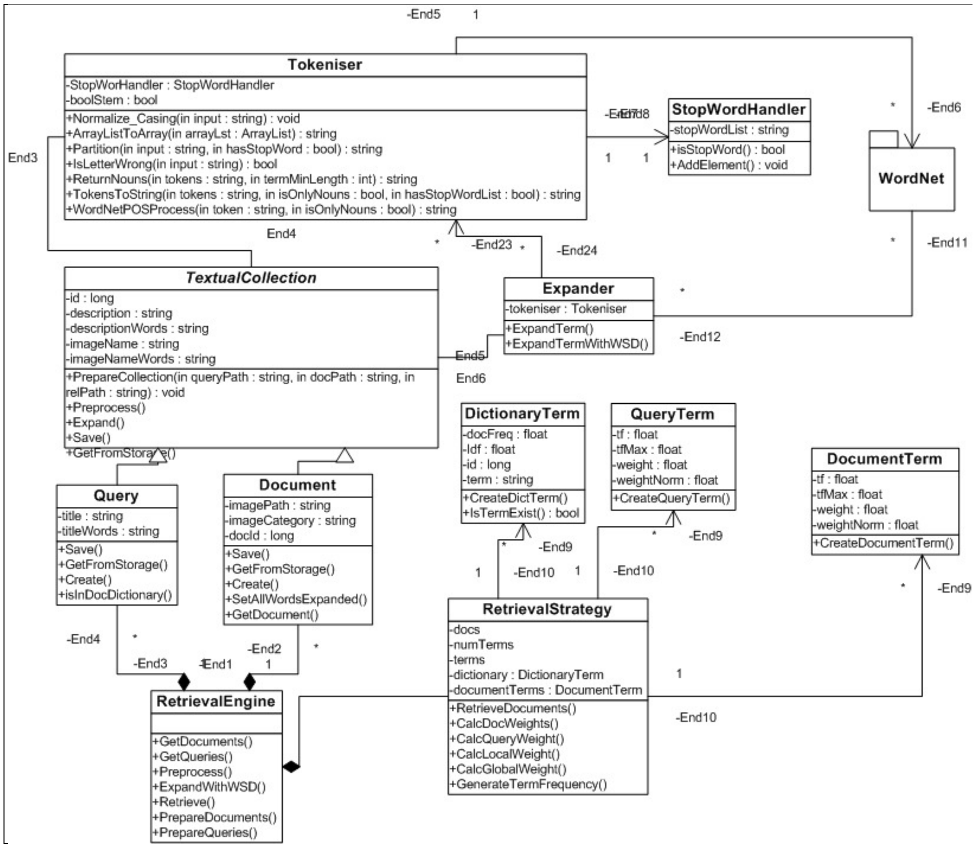


Figure 5.1 UML class diagrams of proposed system.

Then two-level reranking step starts. The first level reranking uses the narrowing-down approach. With the completion of first level, the result set of each query and ranked scores are kept for the second level. The second level is based on cover coefficient (CC) concept. Final similarity score $R(j)$ is calculated using $R_1(j)$ from the first level and query-document similarity score, $R_2(j)$ from the second level. Finally, two-level reranking process is completed and final ranked result sets are generated. Figure 5.2 presents the view of proposed system.

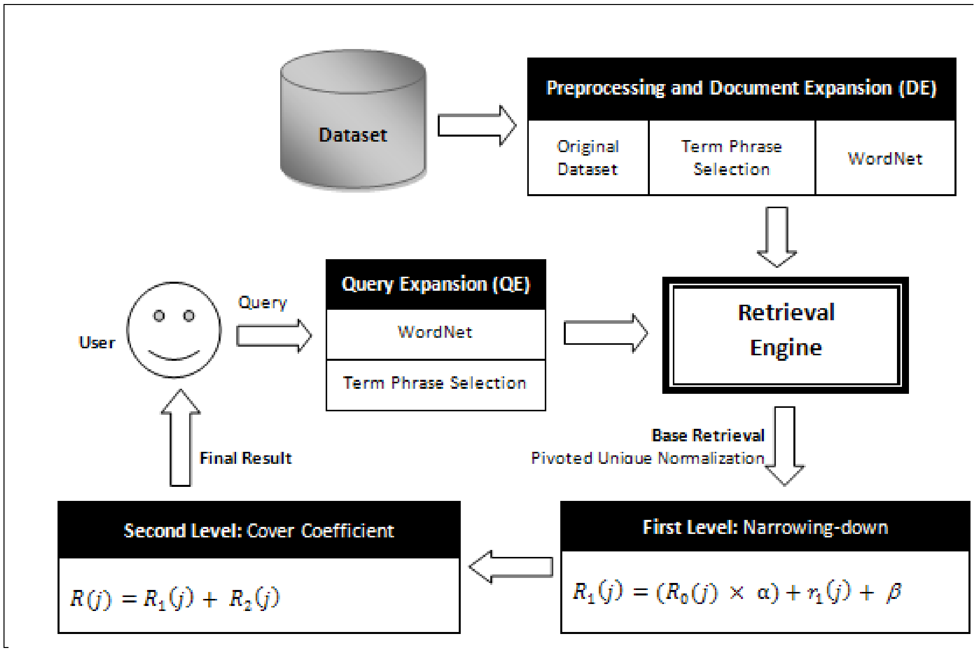

Dataset
Preprocessing and Document Expansion (DE)
Original Dataset
Term Phrase Selection
WordNet
User
Query
Query Expansion (QE)
WordNet
Term Phrase Selection
Retrieval Engine
Final Result
Base Retrieval
Pivoted Unique Normalization
Second Level: Cover Coefficient
$R(j) = R_1(j) + R_2(j)$
First Level: Narrowing-down
$R_1(j) = (R_0(j) \times \alpha) + r_1(j) + \beta$

Figure 5.2 Diagram of document expansion and two-level reranking process.

## 5.3 ImageCLEF WikipediaMM Subtask

ImageCLEF is the cross-language image retrieval track run as part of the Cross Language Evaluation Forum (CLEF) campaign. This track evaluates retrieval of images described by text captions based on queries in a different language; both text and image matching techniques are potentially exploitable. The following tasks:

- a photographic retrieval task,
- a medical retrieval task,
- a robotic image visual task,
- a medical automatic image annotation task, and
- an image retrieval task from a collection of Wikipedia web images.

The whole system has been evaluated with dataset of ImageCLEF's WikipediaMM task (Tsikrika & Kludas, 2009) which provides a test bed for the system-oriented evaluation of visual information retrieval from a collection of Wikipedia images by using both WikipediaMM 2008 and WikipediaMM 2009 topics. The aim is to investigate retrieval approaches in the context of a larger scale and heterogeneous collection of images (similar to those encountered on the Web) that are searched for by users with diverse information needs. It contains approximately 150,000 images that cover diverse topics of interest. These images are associated with unstructured and noisy textual annotations in English.

This is an ad-hoc image retrieval task; the evaluation scenario is thereby similar to the classic TREC ad-hoc retrieval task and the ImageCLEF photo retrieval task: simulation of the situation in which a system knows the set of documents to be searched, but cannot anticipate the particular topic that will be investigated (i.e. topics are not known to the system in advance). The goal of the simulation is: given a textual query (and/or sample images) describing a user's (multimedia) information need, find as many relevant images as possible from the Wikipedia image collection.

Any method can be used to retrieve relevant documents. Concept-based and content-based retrieval methods and, in particular, multimodal approaches that investigate the combination of evidence from different modalities, can be used.

The WikipediaMM task encourages participants to create the topics and perform the relevance assessments themselves. This is similar to the user model followed in INEX, with the difference that participants are not required to get involved in that process. It is an optional step that allows the participants to share in the creation of the test collection.

## 5.4 Dataset and Topics (Queries)

The dataset consists of approximately 150,000 Wikipedia images (in JPEG and PNG formats) provided by Wikipedia users. Each image is associated with user-generated alphanumeric, unstructured metadata in English. Figure 5.3 presents an example image with document number 1027698.

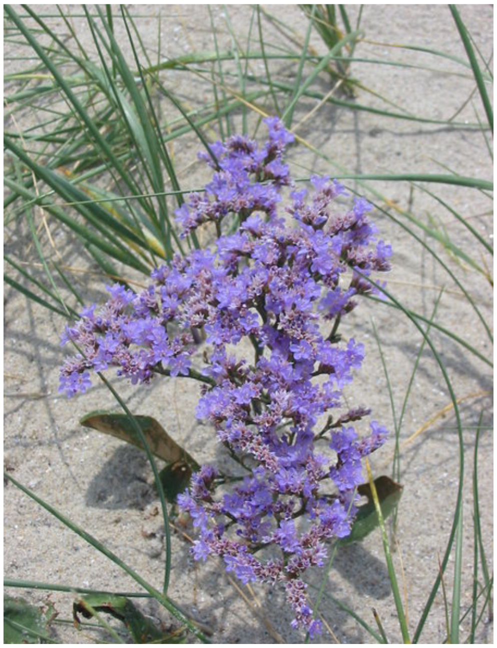


Figure 5.3 Sample image from Wiki dataset with document number 1027698.

Table 5.1 depicts the metadata information of the same image. These metadata usually contain a brief caption or description of the image, the Wikipedia user who uploaded the image, and the copyright information. These descriptions are highly heterogeneous and of varying length.

Table 5.1 XML metadata sample of document, numbered 1027698

**Metadata**

```
<?xml version="1.0"?>
<article>
    <name id="1027698">Sea_lavender.JPG</name>
   <image xmlns:xlink="http://www.w3.org/1999/xlink"
xlink:type="simple"
  xlink:actuate="onLoad"xlink:show="embed"
id="1027698">Sea_lavender.JPG</image>
   <text>
    <wikilink type="internal" parameters="1">
      <wikiparameter number="0" last="1">
         <value>Sea lavender</value>
      </wikiparameter>
    </wikilink> (Limonium vulgare), Picture taken by
    <wikilink type="internal" parameters="1">
      <wikiparameter number="0" last="1">
         <value>User:Donarreiskoffer</value>
      </wikiparameter>
     </wikilink>
  </text>
</article>
```

The topics are multimedia queries that can consist of a textual and a visual part, with the latter being optional. Concepts that might be needed to constrain the results should be added to the title field. Table 5.2 presents a simple topic.

Table 5.2 XML metadata sample of topic, numbered 1

**Topic (Query)**

```
<topic>
  <number> 1 </number>
  <title> cities by night <title>
  <image> http://www.bushland.de/hksky2.jpg </image>
  <narrative> I am decorating my flat and as I like photos of
cities at night, I would like to find some that I could possibly
print into posters. I would like to find photos of skylines or
photos that contain parts of a city at night (including streets
and buildings).Photos of cities (or the earth) from space are not
relevant. </narrative>
</topic>
```

The topics include the following fields. In proposed work, only the textual parts (`<title>`) of topics have been used.

- title: simulates a user who does not have (or does not want to use) example images or other visual information. The query expressed in the topic `<title>` is therefore a text-only query. This profile is likely to fit most users searching digital libraries.
- image: query by one or more images (optional)
- narrative: description of the information need where the definitive definition of relevance and irrelevance are given

We participated into WikipediaMM 2009 subtask to prove the efficiency of proposed work but we have done experiments both using WikipediaMM 2008 and WikipediaMM 2009 topics. The sample topics can be found at the appendix section A.1 and A.2.

### 5.5 Evaluation Measures

In order to evaluate an IR-system, there are several measures used to determine the quality of the system. There are different aspects of an IR-system, all of which contribute to the cumulated quality of the entire system. The different aspects are typically *processing quality,* that is, the time and space efficiency of the system, *search quality*, which is the effectiveness of results and an overall *system quality*, which tries to measure the satisfaction of the user.

*Relevance* can be defined as a measure of how well the result meets the need of the user that issued the query. The purpose of ranking algorithms is to present the user with documents that address their information need. A document is considered to be relevant if it in some way satisfies a user's information need. This does not necessarily mean that the document is useful to a user. As a simple example, if there are two identical documents, both would be considered to be relevant, but in most cases only one of them will be beneficial to the user in resolving their information need – there is no use to see the same document twice.

Saracevic (1999) proposes a hierarchy of relevance. On the lowest level, system or algorithmic relevance is achieved if a document satisfies a query on a syntactic level. More grades of relevance (topical or subject relevance, cognitive relevance or pertinence, situational relevance or utility) determine ever stronger user satisfaction, while motivational or affective relevance completely addresses the user's information need.

The most important two measures are *Precision* and *Recall*. Figure 5.4 shows the relationship between recall, precision, relevant documents (the red line) and the collection (the grey circle).

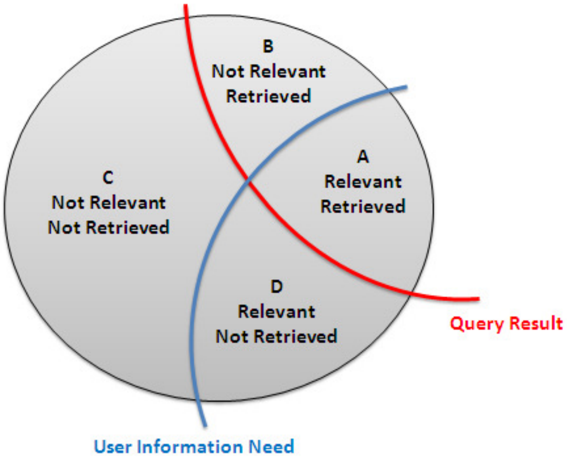


Figure 5.4 Query and user information need presentation.

Precision is the fraction of the relevant documents which has been retrieved:

*Precision* = | *Relevant retrieved* | / | *All retrieved documents* |

*Precision = A / (A+B)*

Consider the example in Table 5.3, where a search engine has retrieved ten documents in response to an imaginary query. Table 5.3 shows a ranked answer set, we can see that five relevant documents are amongst the top ten retrieved documents. Given that there are eight relevant documents in the collection, recall would be |R| = 5/8 = 63%. The recall achieved at cutoff level 10 would be (5/8) 63% in the example, whereas the recall at level 5 would be (2/8) 25%.

Recall is the fraction of the relevant documents which has been retrieved:

*Recall* = | *Relevant retrieved* | / | *All relevant documents* |

*Recall = A / (A+D)*

Precision must always be stated at a certain cutoff level in the case of recall. In the example, precision at the cutoff level of 10 would be Precision@10 = 5/10 = 0.5 = 50%. This is denoted as P@10. Since we know from Table 5.3, that amongst the top 5 results only Document 1 and Document 5 are relevant, we can calculate the precision at 5 documents as P@5 = 2/5 = 40%.

Table 5.3 Sample results with relevance judgments for an imaginary query

| Doc# | Relevancy |
|---|---|
| 1 | Relevant Document |
| 2 | Non-Relevant Document |
| 3 | Non-Relevant Document |
| 4 | Non-Relevant Document |
| 5 | Relevant Document |
| 6 | Relevant Document |
| 7 | Non-Relevant Document |
| 8 | Non-Relevant Document |
| 9 | Relevant Document |
| 10 | Relevant Document |

A standard evaluation strategy for IR-systems is the use of *average precision versus recall* figures. In order to calculate the precision/recall rate as accurate as possible, it is common to run a series of tests with different queries and generate precision/recall figures for all of these. Finally the calculated average with the formula is below;

$$P(r) = \sum_{i=1}^{Nq} \frac{Pi(r)}{Nq} \qquad (5.1)$$

where, $P(r)$ is the average precision at the recall level $r$, $Nq$ is the number of queries used and $Pi(r)$ is the precision at recall level $r$ for the $i$-th query.

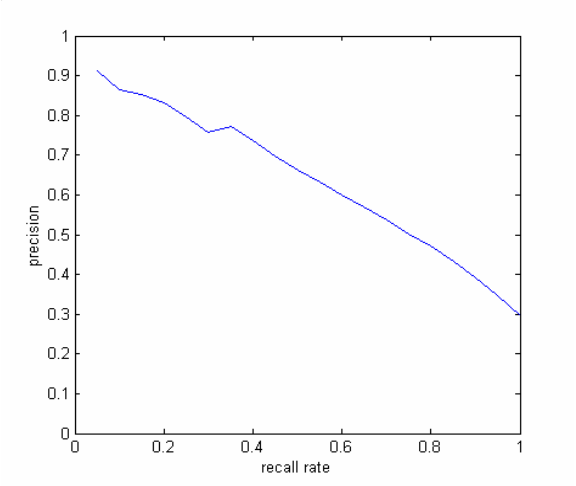


Figure 5.5 Precision versus recall rate graphic.

One of the strengths of this kind of diagrams is that they are an easy way of comparing different retrieval algorithms. The average precision measure is also relatively stable, that is, a change as described above would result in only a fairly small change in the average precision value.

Another advantage is that an average precision figure is influenced to a greater degree by relevant documents at higher ranks, while lower ranked documents play a smaller role. This could be seen as reflecting the perception of the usefulness of a system to a user in a typical retrieval situation. Average precision for the sample results in Table 5.3 is calculated as in equation 5.2.

$$AP\ (Average\ Precision) = \frac{1}{8}\ x\ (\frac{1}{1}+\frac{2}{5}+\frac{3}{6}+\frac{4}{9}+\frac{5}{10}) = 35.5\ \% \qquad (5.2)$$

In this thesis, the proposed system's retrieval performance evaluated using mean average precision (MAP) in addition to P@5 and P@10. MAP is a standard IR evaluation measure, is used to find the mean of the average precisions over a set of queries. For a single information need, Average Precision is the average of the precision value obtained for the set of top $k$ documents existing after each relevant document is retrieved, and this value is then averaged over information needs formulated as in equation 5.3. That is, if the set of relevant documents for an information need *qj* ∈ *Q* is {*d*1, . . . *dmj*} and *Rjk* is the set of ranked Retrieval results from the top result until you get to document *dk*.

$$MAP(Q) = \frac{1}{|Q|}\sum\nolimits_{j=1}^{|Q|}\frac{1}{m_j}\sum_{k=1}^{m_j} Precision(R_k) \qquad (5.3)$$

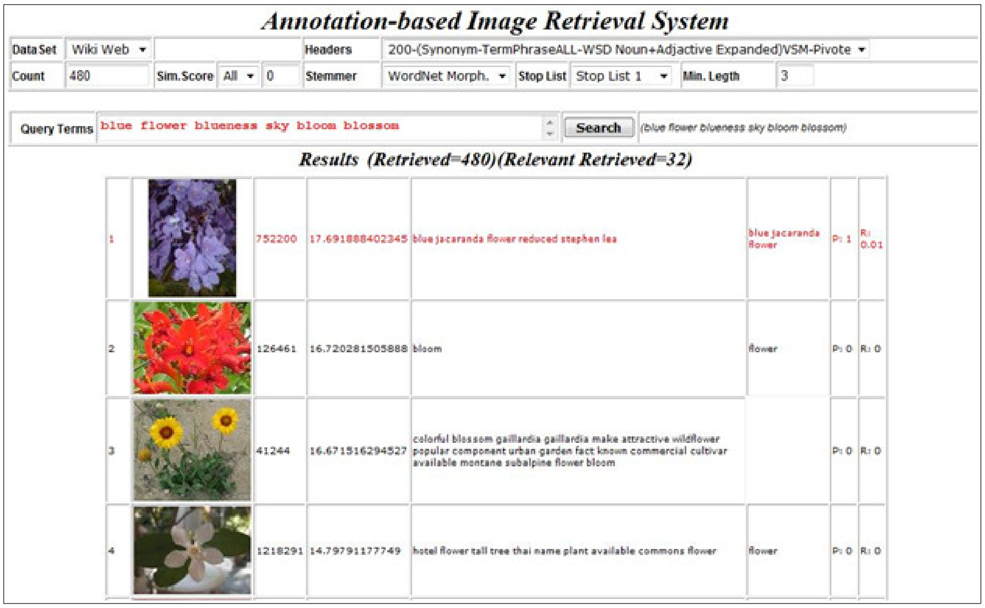
where $Q$ is a set of information needs, each information need $q_j$ has $m_j$ relevant documents, $R_k$ is the set of ranked retrieval results from the top until you get the relevant document *k*.

## 5.6 Development Environment

Microsoft Visual Studio 2008 with C# programming language is used during proposed system's application development cycle. Microsoft Visual Studio is an integrated development environment (IDE) from Microsoft. It can be used to develop console and graphical user interface applications along with Windows Forms applications, web sites, web applications, and web services in both native code together with managed code for all platforms supported by Microsoft Windows, Windows Mobile, Windows CE, .NET Framework and .NET Compact Framework.

Visual Studio includes a code editor supporting IntelliSense as well as code refactoring. The integrated debugger works both as a source-level debugger and a machine-level debugger. Other built-in tools include a forms designer for building GUI applications, web designer, class designer, and database schema designer.

Figure 5.6 presents the developed system's graphical user interface. Some major parameters can be set from top frame of the screen. Dataset parameter has the value Wiki Web by default. In the future, it is easy to adapt proposed system to work with other Web based data sources, like Google Images, Corel Images etc. Header parameter specifies the preprocessed experiment type. Other parameters are used to determine stemming type, stop list tenancy, minimum query length selection and similarity algorithm type. In the middle frame, preprocessed queries can be selected or changed manually by user. Furthermore, evaluation results (precision, recall, MAP) can be displayed, if relevance assessments of queries are introduced to the developed system. Bottom frame shows the retrieved image results in ranked order.



*Annotation-based Image Retrieval System*

Data Set: Wiki Web
Headers: 200-(Synonym-TermPhraseALL-WSD Noun+Adjactive Expanded)VSM-Pivote
Count: 480
Sim.Score: All 0
Stemmer: WordNet Morph.
Stop List: Stop List 1
Min. Legth: 3

Query Terms: blue flower blueness sky bloom blossom

(blue flower blueness sky bloom blossom)

*Results (Retrieved=480)(Relevant Retrieved=32)*

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1 | | 752200 | 17.691888402345 | blue jacaranda flower reduced stephen lea | blue jacaranda flower | P: 1 | R: 0.01 |
| 2 | | 126461 | 16.720281505888 | bloom | flower | P: 0 | R: 0 |
| 3 | | 41244 | 16.671516294527 | colorful blossom gaillardia gaillardia make attractive wildflower popular component urban garden fact known commercial cultivar available montane subalpine flower bloom | | P: 0 | R: 0 |
| 4 | | 1218291 | 14.79791177749 | hotel flower tall tree thai name plant available commons flower | flower | P: 0 | R: 0 |

Figure 5.6 Sample screen from proposed system’s application.

In application, we used WordNet.Net, an open-source .NET Framework library for WordNet developed by Malcolm Crowe and Troy Simpson. It was originally created by Malcolm Crowe and it was known as a C# library for WordNet. It was created for WordNet 1.6 and stayed in its original form until after the release of WordNet 2.0 when Troy gained permission from Malcolm to use the code for freeware dictionary/thesaurus projects. Finally, after WordNet 2.1 was released, Troy released his version of Malcolm's library as an LGPL library known as WordNet.Net (with permission from Princeton and Malcolm Crowe, and in consultation with the Free Software Foundation), which was updated to work with the WordNet 2.1 database.

### 5.7 Proposed System's Database

We used MySQL database to store semi-runs and working results to save time during proposed system's design. MySQL is a relational database management system (RDBMS) that runs as a server providing multi-user access to a number of databases. The MySQL development project has made its source code available under the terms of the GNU General Public License, as well as under a variety of proprietary agreements. MySQL was owned and sponsored by a single for-profit firm, the Swedish company MySQL AB, now owned by Sun Microsystems, a subsidiary of Oracle Corporation. MySQL uses Ranking with Vector Spaces for ordinary full-text queries. Rank, also known as relevance rank, also known as relevance measure, is a number that tells us how good a match is (Garcia, 2006).

Figure 5.7 presents the proposed system's database table structures and relations. All Wiki images are stored in wiki_dr_images table with their dataset identification numbers, category information, image names, file path and description. In addition to this, expanded forms of image descriptions and image names after the preprocess operation, are kept in this table. Docid field is given by system automatically to make retrieval performance better and all table references are done on this field. Wiki_query table stores all queries and related information like query numbers, query titles and query descriptions if exists. Wiki_dictionary table stores dictionary terms with their unique ids, inverse document frequencies and global weights. Wiki_doc_terms table stores term frequency for each document and local weights. Wiki_norm_factor table keeps vector lengths of documents (images). Baseline_wiki_query table stores query terms depending on used expanding technique, if no expanding is done, query terms are stored with their native preprocessed forms.

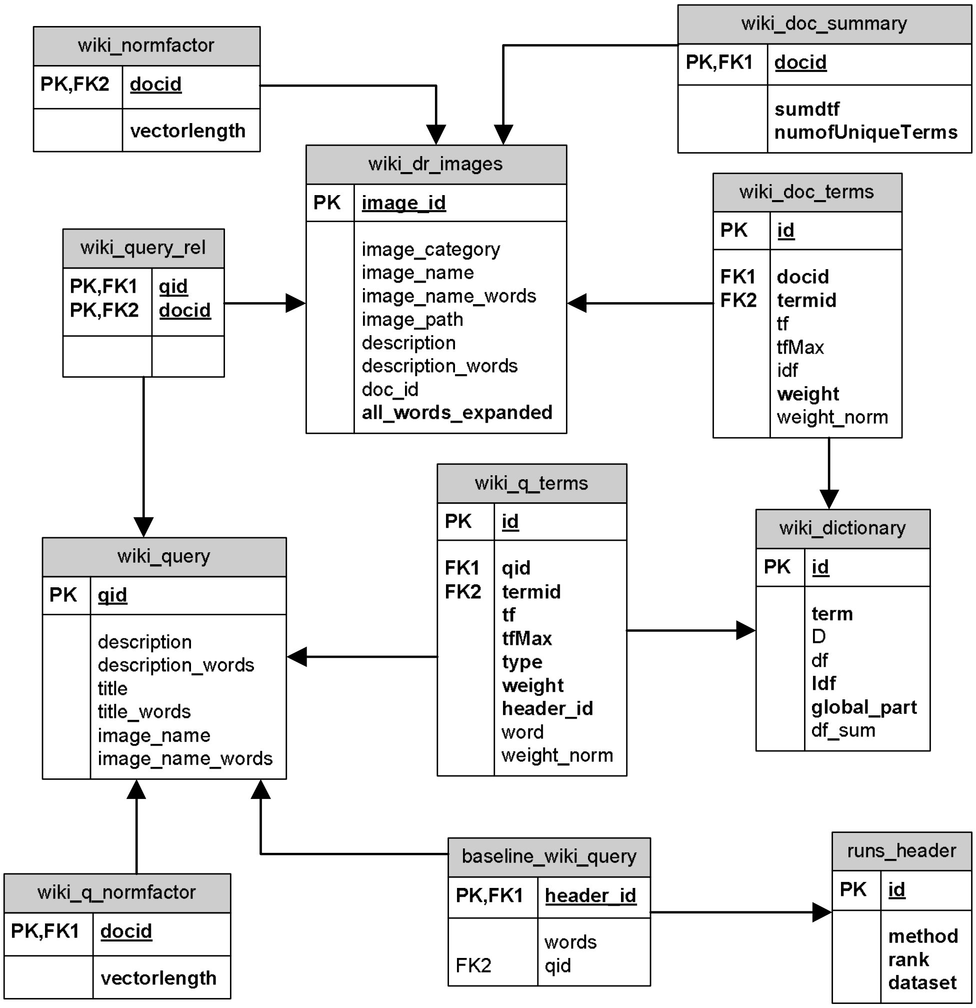


Figure 5.7 Proposed system's database table structures.

## 5.8 Experimental Results

We firstly tested the proposed system with WikipediaMM 2008 queries by using same type of runs, although we did not participate into ImageCLEF 2008. We had the similar promising evaluation results like WikipediaMM 2009. Table 5.4 presents the WikipediaMM 2008 runs and evaluation results.

Table 5.4 Applied techniques and their evaluation results in WikipediaMM 2008 Task

| # | Description | DE | QE (TPS) | QE (WN) | 1st Level Rerank | 2nd Level Rerank | MAP | P@5 | P@10 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | wiki2008_00 | X | | | | | 0.2549 | 0.4453 | 0.3693 |
| 2 | wiki2008_01 | X | X | | | | 0.2555 | 0.4479 | 0.3720 |
| 3 | wiki2008_02 | X | X | X | | | 0.2649 | 0.4537 | 0.3747 |
| 4 | wiki2008_03 | X | X | X | X | | 0.2758 | 0.4587 | 0.3933 |
| 5 | wiki2008_04 | X | X | X | X | | 0.2758 | 0.4587 | 0.3933 |
| 6 | wiki2008_05 | X | X | X | X | X | 0.2765 | 0.4767 | 0.4107 |

Figure 5.8 presents the precision/recall graph of all runs in WikipediMM 2008. We generated the second best MAP and precision results among all participants in WikipediaMM 2008 task of ImageCLEF 2008 contest.

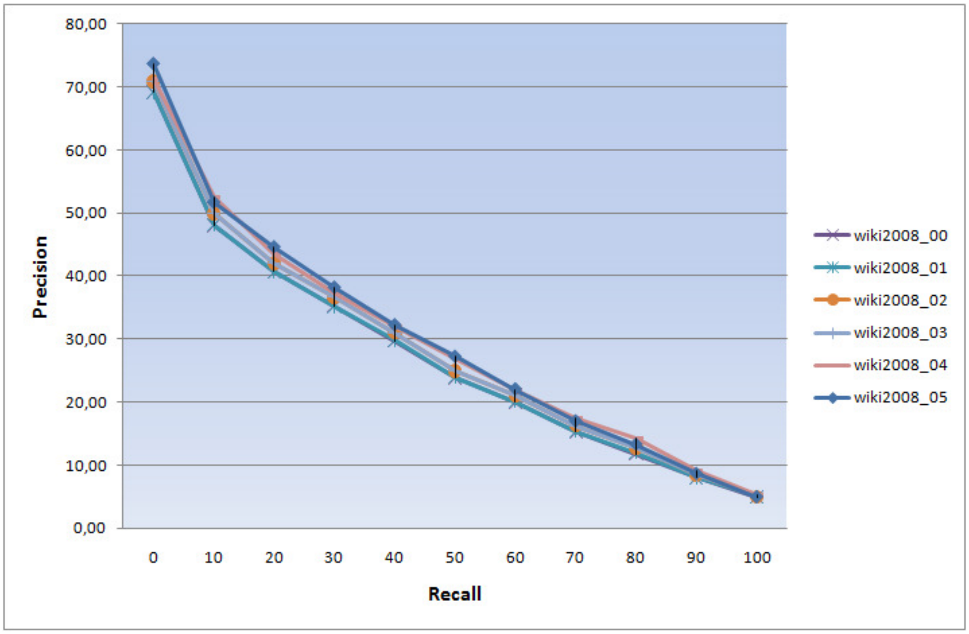


Figure 5.8 Precision/Recall graph of WikipediaMM 2008 experiments.

In WikipediaMM 2009 task, total of 8 groups were participated into and submitted 57 runs; 26 of them text-based retrieval and 31 of also includes content based retrieval. We participated by group name deuceng and conducted six runs. Before the runs, we performed preprocessing and performed aforementioned methods. Besides, we also backup the original forms of documents to calculate the similarity score as a combination of original and expanded dataset. In all runs, we used document expansion and pivoted unique normalization. Few examples from dataset with their original and expanded descriptions are presented in Table 5.5. More examples can be found at appendix section C.1

Table 5.5 Image samples from Wiki dataset

| Id | Image | Description | expanded_form |
|---|---|---|---|
| 10 | 1959ModelPiperPA 24Comanche.jpg | A 1959 model Piper PA 24 Comanche Valleyfield Quebec 2004 | model piper comanche quebec model comanche Comanche member shoshonean people live wyoming mexican border oklahoma bagpiper play bagpipe Quebec french speaking capital province quebec situated saint lawrence river framework hypothetical complex entity process computer program based model circulatory respiratory |
| 1000 217 | Apollo13_splashdown.jpg | Apollo 13 s successful splashdown after a harrowing trip. http://grin.hq.nasa.gov/ABSTRACTS/GPN 2000 001312.htmlNASA s image informationPD USGov NASA | abstract nasa apollo splashdown landing spacecraft sea end space flight smile smiling grinning facial expression characterized turning corner mouth show pleasure amusement NASA independent agency united state government responsible aviation spaceflight journey purpose return trip shopping Apollo Phoebus greek mythology greek god light god poetry music healing son zeus leto twin brother artemis abstraction concept idea associated specific instance loved… |

The differences between the runs were based on the different expansion and reranking techniques. The proposed system's retrieval performance evaluated using mean average precision (MAP), which is described above. Table 5.6 represents the applied techniques for each run and their performance evaluation results.

Table 5.6 Applied techniques and their evaluation results in WikipediaMM 2009 Task

| # | Description | DE | QE (TPS) | QE (WN) | 1st Level Rerank | 2nd Level Rerank | MAP | P@5 | P@10 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | wiki2009_00 | X | | | | | 0.1861 | 0.3244 | 0.2956 |
| 2 | wiki2009_01 | X | X | | | | 0.1865 | 0.3422 | 0.2978 |
| 3 | wiki2009_02 | X | X | X | | | **0.2358** | **0.4844** | 0.3933 |
| 4 | wiki2009_03 | X | X | X | X | | **0.2375** | **0.4933** | 0.4000 |
| 5 | wiki2009_04 | X | X | X | X | | **0.2375** | **0.4933** | 0.4000 |
| 6 | wiki2009_05 | X | X | X | X | X | **0.2397** | **0.5156** | 0.4000 |

The first run (wiki2009_00) is the base retrieval, in which any query expansion and reranking technique is built upon it. The MAP and P@5 values are 0.1861 and 0.3244, respectively. Top 20 first run (wiki2009_00) results for query 113, *baby*, are presented in table 5.7. More query examples as the first run result can be found at appendix section D.

Table 5.7 Top 20 result sets of run “wiki2009_00” for query 113

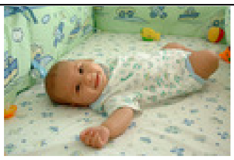
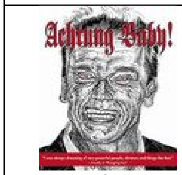
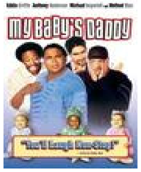
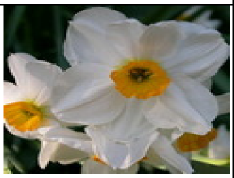
| | | | |
|---|---|---|---|
| <br>**1**<br>**575409.jpeg:** flower baby blue eyes nemophila var photograph point reyes national | <br>**2**<br>**1050468.jpeg:** browse view stock smiling baby lying soft cot furniture baby | <br>**3**<br>**257433.jpeg:** commons baby emu baby emu beacon hill park victoria british columbia Canada | <br>**4**<br>**246254.jpeg:** baby fetal alcohol syndrome come baby |
| <br>**5**<br>**207539.jpeg:** jerry baby file | <br>**6**<br>**193486.jpeg:** photograph geographer baby white fir mount whitney | <br>**7**<br>**1218291.jpeg:** hotel flower tall tree thai name plant available commons flower | <br>**8**<br>**208546.jpeg:** baby week age captive breed crested gecko |
|  <br>**9**<br>**25675.jpeg:** baby poster depicting california celebrity governor arnold | <br>**10**<br>**1031633.jpeg:** victoria flower see victoria caption flower volleyball hot .. | <br>**11**<br>**156731.jpeg:** henson baby title screen film | <br>**12**<br>**45864.jpeg:** rosemary baby dvd cover |

Table 5.7 Top 20 result sets of run “wiki2009_00” for query 113 (Continue…)

| | | | |
|---|---|---|---|
| <br> **13** <br> **249325.jpeg:** baby bear | <br> **14** <br> **97506.jpeg:** chipmunk series character baby film |  <br> **15** <br> **87578.jpeg:** dvd case cover baby daddy deem fair | <br> **16** <br> **228796.jpeg:** guinea pig baby hours old |
| <br> **17** <br> **77172.jpeg:** baby sleeping robert | <br> **18** <br> **116002.jpeg:** unknown flower adam help identify flower form narcissus … |  <br> **19** <br> **212885.jpeg:** unknown flower adam help identify flower form narcissus .. | <br> **20** <br> **146342.jpeg:** screen shot million dollar baby film night category |

The second run (wiki2009_01) includes query expansion using only TPS (Term Phrase Selection). The MAP and top precision values of the second run are slightly better than base retrieval.

In the third run (wiki2009_02), we expanded the queries using both TPS and WordNet with the same document expansion approaches. The third run’s MAP and P@5 values are 0.2358 and 0.4844, respectively. The experiment result shows that the run performs considerably better since our proposed expansion techniques for documents and queries are same. Measurement metrics of query 113 results for wiki2009_00 and wiki2009_02 are presented in Table 5.8.

Table 5.8 Measurement metrics of query 113 results

| Metrics | wiki2009_00 | wiki2009_02 |
|---|---:|---:|
| MAP | 0.2844 | 0.3469 |
| P@5 | 0.6000 | 0.8000 |
| P@10 | 0.3000 | 0.4000 |
| P@15 | 0.3333 | 0.4000 |
| P@20 | 0.2500 | 0.3000 |
| P@30 | 0.2000 | 0.2667 |
| P@100 | 0.1600 | 0.1700 |
| P@200 | 0.0950 | 0.1000 |
| P@500 | 0.0400 | 0.0400 |
| P@1000 | 0.0200 | 0.0200 |

For example P@5 is improved from 0.6 to 0.8 that means the number of relevant images is increased from 3 to 4 on top 5 results, and also P@10 is improved from 0.3 to 0.4 that means the number of relevant images is increased from 3 to 4 on top 10 results. These improvements for expanded query 113, *baby infant young child*, can be seen in table 5.9.

Table 5.9 Top 20 result sets of run "wiki2009_02" for query 113

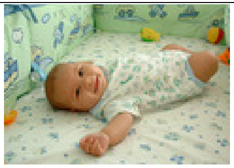
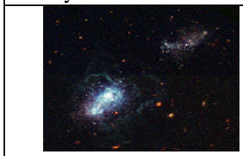
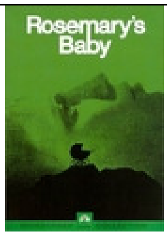
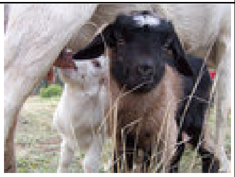
| | | | |
|---|---|---|---|
| <br>**1**<br>**1050468.jpeg:** browse view stock smiling baby lying soft cot furniture baby | <br>**2**<br>**239034.jpeg:** net entitled net gallery baby feeding copy Michael | <br>**3**<br>**207539.jpeg:** jerry baby file | <br>**4**<br>**246254.jpeg:** baby fetal alcohol syndrome come baby |
| <br>**5**<br>**1608990.jpeg:** hubble baby galaxy grown universe archive release archive release nasa hubble baby | <br>**6**<br>**208546.jpeg:** baby week age captive breed crested gecko | <br>**7**<br>**575409.jpeg:** flower baby blue eyes nemophila var photograph .. | <br>**8**<br>**193486.jpeg:** photograph geographer baby white fir mount whitney |

Table 5.9 Top 20 result sets of run “wiki2009_02” for query 113 (Continue…)

| | | | |
|---|---|---|---|
| <br>**9**<br>**257433.jpeg:** commons baby emu baby emu beacon hill … | Achtung Baby!<br>**10**<br>**25675.jpeg:** baby poster depicting california celebrity governor arnold |  <br>**11**<br>**45864.jpeg:** rosemary baby dvd cover | <br>**12**<br>**116448.jpeg:** young zucchini |
| <br>**13**<br>**249325.jpeg:** baby bear | <br>**14**<br>**77172.jpeg:** baby sleeping Robert | Jim Henson's Muppet Babies<br>**15**<br>**156731.jpeg:** henson baby title screen film | <br>**16**<br>**228796.jpeg:** guinea pig baby hours old |
| <br>**17**<br>**97506.jpeg:** chipmunk series character baby film | <br>**18**<br>**225760.jpeg:** baby goat | my baby's daddy<br>**19**<br>**87578.jpeg:** dvd case cover baby daddy deem fair | <br>**20**<br>**251123.jpeg:** baby |

The next three runs show that our reranking approach provides an increase in precision. In the fourth run (wiki2009_03), we conducted first-level reranking named narrowing-down approach, including re-indexing. The experiment results are slightly better again, especially with the impact of $\beta$ parameter. The MAP and P@5 values are 0.2375 and 0.4933, respectively.

The difference between the fifth run (wiki2009_04) and the fourth run (wiki2009_03) is the documents in the result set above a threshold ranking value are used for the first level reranking process, but the experiment results are same. Final run, wiki2009_05 (wiki2009_05), includes additional CC based second level reranking approach over the result set of the fifth run.

As can be seen from the MAP values in table, the best experiment results obtained from the sixth run (wiki2009_05). The MAP and P@5 values of the sixth run are 0.2397 and 0.5156, respectively. Figure 5.9 presents the precision/recall graph of all runs in WikipediMM 2009. We generated the best MAP and precision results among all participants in WikipediaMM 2009 task of ImageCLEF 2009 contest. WikipediaMM 2009's evaluation results can be found at appendix section B.2.

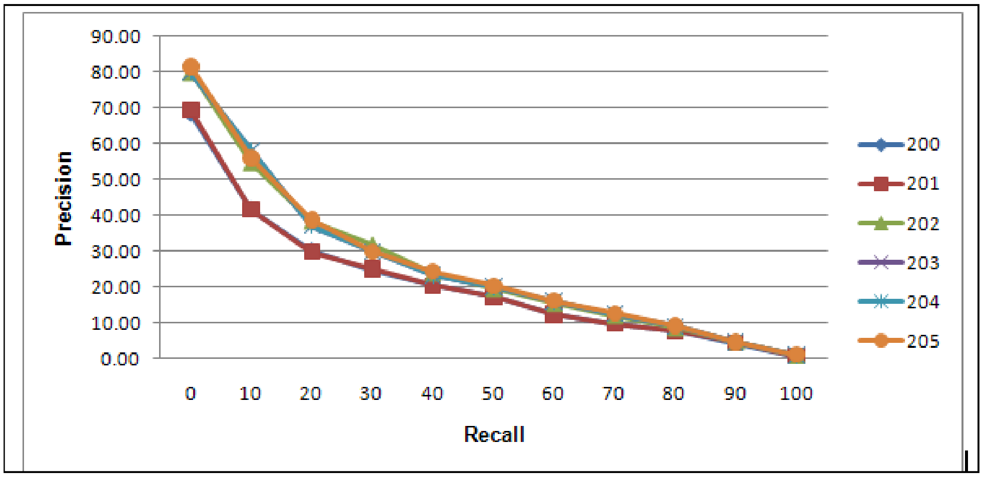


Figure 5.9 Precision/Recall graph of WikipediaMM 2009 experiments.

# CHAPTER SIX
# CONCLUSION

## 6.1 Conclusion

In this thesis we presented an ABIR system using new expansion technique for both documents and queries (WordNet, WSD, similarity functions) which increases recall but decreases precision. So, we applied a two-level reranking approach to increase precision. The proposed system has the following contribution to the field:

i. A new expansion approach used.

   a. Same expansion method used for both documents and queries. Text retrieval community studied query expansion extensively. However, in literature, document expansion has not been thoroughly researched for information retrieval. Expanding the poorly defined documents by adding new terms may result in higher ranking performance. Similarly, expanding the queries and widening the search terms increase the recall value by bringing more relevant documents which is not matching literally with the original query.

   b. Since we have strong expansion approaches, there is no need to human interventions such as relevance feedback. Also in our experiments we didn’t make any manual contribution or annotation to increase the performance.

ii. A new two-level reranking approach used.

   a. Document expansion generally results a high recall with low precision. Thus, we proposed a two level reranking approach to move relevant documents upward. The first level of reranking

b. approach forms a narrowing-down phase and includes re-indexing. In other words, we filter out non-relevant documents based on initial similarity scores and this operation drastically reduces both the number of documents and the number of terms. It is a simple but powerful approach. The second level of reranking method is cover coefficient based. Although, C3M originally developed for text clustering, the presented novel approach shows that, C3M can also be used for a reranking step in ABIR.

iii. The proposed system does not have boundaries, and hence can be extended with other techniques such as relevance feedback, to use their benefits.

iv. It is easy to integrate proposed system to retrieve other Web based image collections (e.g., Google images, Corel images etc.).

In order to evaluate our approaches we have participated ImageCLEF2009 WikipediaMM subtask. The proposed system was evaluated using the huge Wiki dataset which approximately contains 150.000 images with their annotations. We used 120 queries; 45 of them from WikiMM2008, 75 of them from WikiMM2009. Experimentation showed that our suggestion to annotation based image retrieval system is promising, so that it generated the best MAP and precision results among all participants in WikipediaMM task of ImageCLEF 2009 contest. We also experimented with ImageCLEF2008 WikipediaMM queries and had the similar promising results. The results also showed that document expansion, effective term selection to annotations and two-level reranking plays an important role in text-based image retrieval.

### 6.2 Future Works

For the further studies, the remaining number of open issues, each of requires an individual research are showed below;

- First level reranking, narrowing-down approach, can be extended.

- New semi-supervised reranking techniques, like document categorization, probabilistic weighting, collaborative filtering or combination of them can be used.

- As we told in the previous subsection, the proposed system can be experimented using other Web based image collections (e.g., Google images, Corel images etc.)

- As we told in the previous subsection, the proposed approach can be extended with many other image retrieval techniques such as, relevance feedback.

- The proposed solution can be experimented with multi lingual datasets, but WordNet must be changed with a new approach.

- The proposed solution can lead to new researches including semantic web, semantic indexing, and development of image ontology automatically.

As we discussed, there are open research issues that are not covered in this thesis and they need to be investigated.

## REFERENCES

Allan, J., Leuski, A., Swan, R., & Byrd, D. (2001). Evaluating combinations of ranked lists and visualizations of inter-document similarity. *Information Processing & Management* , 37(3):435-458.

Balanski, J., & Danilowicz, C. (2005). Re-ranking method based on inter-document distances. *Information Processing & Management* , 41(4):759-775.

Banerjee, S., & Pederson, T. (2003). *An adapted Lesk algorithm for word sense disambiguation using WordNet.* Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing, (pp. 136-145). London.

Billerbeck, B., & Zobel, J. (2005). *Document expansion versus query expansion for ad hoc retrieval.* The 10th Australasian Document Computing Symposium, (pp. 34-41). Sydney, Australia.

Bookstein, A. (1982). *Explanation and generalization of vector models in information retrieval,* in G. Salton and H.-J. Schneider, eds, "Proc. ACM-SIGIR Int. Conf. on Research and Development in Information Retrieval". *Springer-Verlag* , 118-132.

Buckley, C. (1993). The importance of proper weighting methods. *Human Language Technology Conference* (pp. 349-352). Princeton, New Jersey: Association for Computational Linguistics.

Callan, J., Croft, W. B., & Harding, S. M. (1992). *The inquery retrieval system.* In Proceedings of the Third International Conference on Database and Expert Systems Applications (pp. 78-83). Springer-Verlag.

Can, F., & Ozkarahan, E. A. (1990). Concepts and Effectiveness of the Cover Coefficient Based Clustering Methodology for Text Databases. *ACM Transactions on Database Systems*, *15*, p. 4.

Carbonell, J., & Goldstein, J. (1998). *The use of mmr, diversity-based reranking for reordering documents and producing summaries.* In SIGIR ’98 Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval (pp. 335-336). New York: ACM.

Carson, C., Thomas, M., Belongie, S., Hellerstein, J. M., & Malik, J. (1999). *Blobworld: A system for region-based image indexing and retrieval.* Third International Conference on Visual Information Systems.

Carvalho, V. R., Elsas, J. L., Cohen, W. W., & Carbonell, J. G. (2008). *A meta learning approach for robust rank learning.* Proceedings of Learning to Rank for Information Retrieval Workshop (pp. 15-23). ACM: ACM.

Chang, N. S., & Fu, K. S. (1979). *A relational database system for images.* Technical Report TR-EE, Purdue University.

Chang, N. S., & Fu, K. S. (1980). Query-by-Pictorial-Example. *IEEE Trans. Software Eng. , 6*, 519-524.

Chang, S. K., & Kunii, T. L. (1981). Pictorial Data-Base Systems. *Computer , 14*, 13-21.

Choi, Y., & Rasmussen, E. M. (2002). Users' relevance criteria in image retrieval in American history. *Information Processing & Management , 38* (5), 695-726.

Cristianini, N., Shawe-Taylor, J., & Lodhi, H. (2002). "Latent semantic kernels". *J. Intell. Inf. Syst , 18* (2-3), 127-152.

Efthimiadis, E. N. (1996). Query expansion. *Annual Review of Information Science and Technology , 31*, 121-187.

El Kwae, E. A., & Kabuka, M. R. (2000). *Efficient Content-Based Indexing of Large Image Databases.* ACM Trans. Information Systems,  pp. 171-210.

Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Huang, Q., Dom, B., et al. (1995). Query by Image and Video Content: The QBIC System. *Computer , 28* (9), 23-32.

Foskett, D. J. (1997). Thesaurus. In K. S. Jones, & P. Willet, *Readings in Information Retrieval* (pp. 111-134). San Francisco, CA: Morgan Kaufman.

Frankel, C., Swain, M. J., & Athitsos, V. (1996). *Webseer: An image search engine for the world wide web.* Technical Report TR-96-14, University of Chicago, Computer Science Department.

Furnas, G., Landauer, T. K., Gomez, L. M., & Dumais, S. T. (1987). "The vocabulary problem in human-system communication". *Communications of the ACM , 30* (11), 964-971.

Garcia, E. (2006). *Implementation and application of term weights in mysql environment.* Retrieved 10 2008, from http://www.miislita.com/term-vector/term-vector-5-mysql.html

Grossman, D., & Frieder, O. (2004). In *Information Retrieval: Algorithms and Heuristics (2nd Edition)* (p. 332). Dordrecht: Springer Publishers.

Hughes, A., Wilkens, T., Wildemuth, B., & Marchionini, G. (2003). *Text or pictures? an eyetracking study of how people view digital video surrogates.* Proceedings of the International Conference on Image and Video Retrieval, (pp. 271-280).

Ide, E., & Salton, G. (1971). Interactive search strategies and dynamic file organization in information retrieval, in G. Salton, ed. In *"The SMART Retrieval System – Experiments in Automatic Document Processing"* (pp. 373-393). Englewood Cliffs, NJ: Prentice-Hall.

Jeon, J., Lavrenko, V., & Manmatha, R. (2003). *Automatic image annotation and retrieval using cross-media relevance models.* Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, (pp. 119-126).

Lee, K., Park, Y., & Choi, K. (2001). Re-ranking model based on document clusters. *Information Processing & Management* , 37(1):1-14.

Lesk, M. (1986). *Automatic sense disambiguation using machine readable dictionaries: how to tell a pine code from an ice cream cone.* Proceedings of the 5th annual international conference on Systems documentation (pp. 24-26). ACM Press.

Lew, M. S., Lempinen, K., & Huijsmans, D. P. (1997). *Webcrawling using sketches.* Technical report, Leiden University, Computer Science Department, The Netherlands.

Li, Y. C., & Meng, H. M. (2003). Document expansion using a side collection for monolingual and cross-language spoken document retrieval. *"ISCA Workshop on Multilingual Spoken Document"*, (pp. 85-90). Hong Kong.

Li, Y., Bandar, Z. A., & McLean, D. (2003). *An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources.* IEEE Trans. On Knowledge and Data Engineering, (pp. 15(4):871–882).

Lingpeng, Y., Donghong, J., Guodong, Z., & Yu, N. (2005). Improving retrieval effectiveness by using key terms in top retrieved documents. *Advances in Information Retrieval* , 169-184.

Mandala, R., Tokunaga, T., Tanaka, H., Okumara, A., & Satoh, K. (1998). Ad hoc retrieval experiments using WordNet and automatically constructed thesauri. In E. M. Voorhees, & D. K. Harman, *"Proc. Text Retrieval Conf. (TREC)"* (pp. 475-480). Gaithersburg, MD: National Institute of Standards and Technology Special Publication.

Manning, C. D., Raghavan, P., & Schütze, H. (2009). *An Introduction to Information Retrieval.* Cambridge University Press.

Metzler, D., & Kanungo, T. (2008). *Machine learned sentence selection strategies for query based summarization.* Proceedings of Learning to Rank for Information Retrieval Workshop (pp. 40-47). ACM: ACM.

Miller, G. A. (1990). Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography , 3*, 235-312.

Ogle, V. E., & Stonebraker, M. (1995). Chabot: Retrieval from a Relational Database of Images. *Computer , 28*, 40-48.

Park, G., Baek, Y., & Lee, H.-K. (2005). Reranking algorithm using post retrieval clustering for content based image retrieval. *Information Processing & Management , 41* (2), 177-194.

Qiu, & Frei, H. P. (1993). *Concept based query expansion.* Proceedings of ACM SIGIR'93 Conference on Research and Development in Information Retrieval, (pp. 160-169). Pittsburgh, USA.

Resnik, O. (1999). Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity and Natural Language. *Journal of Artificial Intelligence Research* , 11:95–130.

Richardson, R., Smeaton, A., & Murphy, J. (1994). *Using WordNet as a Knowledge Base for Measuring Semantic Similarity Between Words.* Techn. Report Working paper CA-1294, School of Computer Applications, Dublin City University, Dublin, Ireland.

Rigutini, L., Papini, T., Maggini, M., & Scarselli, F. (2008). *Learning to rank by neural-based sorting algorithm.* Proceedings of Learning To Rank for Information Retrieval Workshop (pp. 1-8). ACM: ACM.

Salton, G. (1971). *The SMART Retrieval System – Experiments in Automatic Document Processing*. NJ, Englewood Cliffs: Prentice-Hall.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for information retrieval. *Journal of the American Society for Information Science* , 18(11):613-620.

Santini, S., & Jain, R. (2000). Integrated Browsing and Querying of Image Databases. *IEEE Multimedia*, *7*, pp. 26-39.

Saracevic, T. (1999). Information science. *Journal of the American Society for Information Science , 50* (12), 1051-1063.

Scholer, F., Williams, H. E., & Turpin, A. (2004). "Query association surrogates for web search". *Journal of the American Society for Information Science and Technology , 55* (7), 637-650.

Singhal, A., & Pereira, F. (1999). *Document Expansion for Speech Retrieval.* In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, (pp. 34-41). Berkeley, California, USA.

Smeaton, A. F. (1995). *TREC-4 Experiments at Dublin City University: Thresholding Posting Lists, Query Expansion with WordNet, and POS Tagging of Spanish.* Proceedings of TREC-4 Conference, (pp. 373-390). Gaithersburg, USA.

Smith, J. R. (1997). *Integrated Spatial and Feature Image Systems: Retrieval, Compression and Analysis.* PhD. thesis, Graduate School of Arts and Sciences,Columbia University.

Smith, J. R., & Chang, S. F. (1996). Querying by Color regions using the VisualSeek Content-Based Visual Query System. *Intelligent Multimedia Information.* IJCAI'96.

Tsikrika, T., & Kludas, J. (2009). *Overview of the wikipediaMM task at ImageCLEF 2009.* CLEF working notes 2009, Corfu, Greece.

Tversky, A. (1977). Features of Similarity. *Psychological Review* , 84(4):327–352.

van Rijsbergen, C. J. (2000). "Another look at the logical uncertainty principle". *Kluwer International Journal of Information Retrieval , 2* (1), 77-91.

van Rijsbergen, C. J. (1979). *Information Retrieval.* University of Glasgow, Dept. of Computer Science.

Voorhees, E. M. (1994). *Query expansion using lexical-semantic relations.* In W. B. Croft, & C. J. Rijsbergen, "Proc. ACM-SIGIR Int. Conf. on Research and Development in Information Retrieval" (pp. 61-69). Dublin, Ireland: Springer-Verlag.

Wang, J. Z., Li., J., & Wiederhold, G. (2001). SIMPLIcity, Semantics-Sensitive Integrated Matching for Picture Libraries. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, *vol. 23(9).*

Wong, S. K., & Yao, Y. Y. (1995). On Modeling Information Retrieval with Probabilistic Inference. *ACM Trans. Information Systems*, *13*, pp. 38-68.

Wu, J. K. (1997). Content-Based Indexing of Multimedia Databases. *IEEE Trans. Knowledge and Data Eng.*, *9*, pp. 978-989.

Wu, Q., Iyengar, S. S., & Zhu, M. Web Image Retrieval Using Self-Organizing Feature Map. *J. Am. Soc. Information Science and Technology*, *52*, pp. 868-875. 2001.

Wu, Z., & Palmer, M. (1994). *Verb Semantics and Lexical Selection.* Annual Meeting of the Associations for Computational Linguistics, (pp. 133-138). Las Cruces, New Mexico.

Zhou, T. S. (2002). Unifying keywords and visual contents in image retrieval. *IEEE Multimedia , 9* (2), 23-33.

# APPENDIX A

## A.1 Wiki 2008 queries (5 of 75)

| Qid | description | title |
|---|---|---|
| 1 | I took a picture of a beautiful blue flower image and I would like some more similar images. So I m looking for images of blue flowers. It doesn t have to be exactly the same sort of flower, but other flowers of other colours are not relevant. | blue flower |
| 2 | Among the natural views, I like sunsets by the sea. I m trying to make a collection of such beautiful views; therefore, I seek images of sea sunsets. A relevant image is one that contains both the sea and the sunset. | sea sunset |
| 3 | Find images of a red ferrari. Even though I can not afford one, I like to watch them. Relevant pictures show a red ferrari. | ferrari red |
| 4 | When I was young my family had lots of animals, but my brother s favorite was a cute white cat named "Snow White". Tomorrow my brother is getting married and I am preparing a slideshow about the most remarcable events of his life. Unfortunately, I have no pictures of Snow White. Thus, I am looking for a picture of a cat which is as white as possible. As long as the cat is mostly white, if possible on its back, that should be fine. | white cat |
| 5 | My father is fan of race cars. For his birthday, I want to make an album representing different pictures of silver race cars. The rallying or formula cars should be on the race track or in the pit stop. It could also be possible to have several cars in the picture, if the focus is on a single, silver car. | silver race car |

## A.2 Wiki 2009 queries (5 of 45)

| qid | Title |
|---|---|
| 76 | Shopping in a market |
| 77 | real rainbow |
| 78 | sculpture of an animal |
| 79 | stamp without human face |
| 80 | orthodox icons with Jesus |

## A.3 Generated Term-phrases (87 of 6,808)

| Term phrase | Term phrase | Term phrase |
|---|---|---|
| command module | rhythm section | arrested development |
| heat flash | air medal | professional wrestling |
| electrical switch | lunar module | australian state |

A.3 Generated Term-phrases (Continue…)

| | | |
|---|---|---|
| edward white | james bond | water tower |
| south America | world war | science fiction |
| vicar apostolic | railway station | visible light |
| saint joseph | great hall | adam smith |
| road runner | forbidden city | slow motion |
| office furniture | colonel blimp | digital camera |
| park avenue | web site | alfred Tennyson |
| amazon river | limited edition | tin plate |
| bob Dylan | riot gun | file name |
| comic strip | web browser | adjutant general |
| albert Einstein | bank note | mississippi river |
| central bank | red bay | natural history |
| united kingdom | saint Thomas | west side |
| film festival | ten thousand | county council |
| coal black | balmoral castle | ice cream |
| art history | great Britain | world bank |
| old style | key west | television show |
| video game | football league | nelson Mandela |
| european union | british Columbia | cape town |
| historical record | air force | south Africa |
| stamp collection | fishing boat | political prisoner |
| earth science | cargo ship | life sentence |
| underground railroad | prince albert | south African |
| air defense | federal soldier | nobel prize |
| van eyck | south side | coconut palm |
| european community | soviet union | |

# APPENDIX B

## B.1 Wiki 2009 Evaluation Summary Results

| | deuwiki2009_200 | deuwiki2009_201 | deuwiki2009_202 | deuwiki2009_203 | deuwiki2009_204 | deuwiki2009_205 |
|---|---|---|---|---|---|---|
| **P5** | 0,3244 | 0,3422 | 0,4844 | 0,4933 | 0,4933 | 0,5156 |
| **P10** | 0,2956 | 0,2978 | 0,3933 | 0,4 | 0,4 | 0,4 |
| **P15** | 0,2578 | 0,2607 | 0,3437 | 0,3437 | 0,3437 | 0,3422 |
| **P20** | 0,2333 | 0,2389 | 0,3189 | 0,3111 | 0,3111 | 0,3133 |
| **P30** | 0,2037 | 0,2059 | 0,2689 | 0,2674 | 0,2674 | 0,2696 |
| **P100** | 0,1304 | 0,1316 | 0,1507 | 0,1522 | 0,1522 | 0,1527 |
| **P200** | 0,0924 | 0,0927 | 0,1002 | 0,102 | 0,102 | 0,1024 |
| **P500** | 0,0508 | 0,0512 | 0,0547 | 0,0554 | 0,0554 | 0,0558 |
| **P1000** | 0,0285 | 0,0285 | 0,03 | 0,03 | 0,03 | 0,03 |

### B.2 Evaluation Results for All Submitted Runs in the WikipediaMM 2009 Ranked by MAP

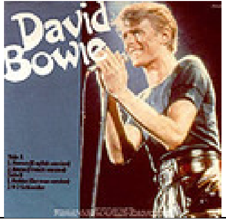
| | Partic. | Run | Modality | Topic field(s) | FB/QE | MAP | P@10 | P@20 | R-prec. | Number of retrieved docs | Number of relevant retrieved docs |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | deuceng | wiki2009_05 | TXT | TITLE | QE | 0.2397 | 0.4000 | 0.3133 | 0.2683 | 43052 | 1351 |
| 2 | deuceng | wiki2009_04 | TXT | TITLE | QE | 0.2375 | 0.4000 | 0.3111 | 0.2692 | 39257 | 1351 |
| 3 | deuceng | wiki2009_03 | TXT | TITLE | QE | 0.2375 | 0.4000 | 0.3111 | 0.2692 | 43052 | 1351 |
| 4 | deuceng | wiki2009_02 | TXT | TITLE | QE | 0.2358 | 0.3933 | 0.3189 | 0.2708 | 43052 | 1352 |
| 5 | … | … | … | … | … | … | … | … | … | … | … |
| 15 | … | … | … | … | … | … | … | … | … | … | … |
| 16 | … | … | … | … | … | … | … | … | … | … | … |
| 17 | … | … | … | … | … | … | … | … | … | … | … |
| 18 | deuceng | wiki2009_01 | TXT | TITLE | QE | 0.1865 | 0.2978 | 0.2389 | 0.2146 | 41242 | 1283 |
| 19 | deuceng | wiki2009_00 | TXT | TITLE | NOFB | 0.1861 | 0.2956 | 0.2333 | 0.2133 | 41242 | 1283 |
| 57 | … | … | … | … | … | … | … | … | … | … | … |

# APPENDIX C

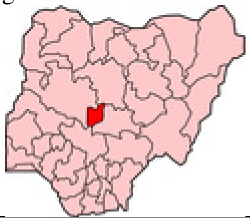
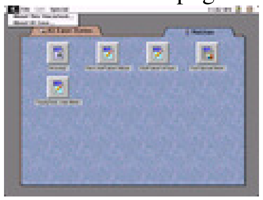
## C.1 Document Expansion Samples (10.jpg, 1000217.jpg, 100008.jpg, 1000213.jpg, 100002.jpg, 100059.jpg, 10005.png, 100051.png)

| Id | Image | image_term | Description | description_term | expanded_terms |
|---|---|---|---|---|---|
| 10 | 1959ModelPiperPA 24Comanche.jpg<br> | model comanche | A 1959 model Piper PA 24 Comanche Valleyfield  Quebec 2004 | model piper comanche quebec | model piper comanche quebec  model comanche Comanche  member shoshonean people live wyoming mexican border oklahoma bagpiper  play bagpipe Quebec  french speaking capital province quebec situated saint lawrence river framework  hypothetical complex entity process computer program based model circulatory respiratory |
| 1000217 | Apollo13_splashdown.jpg<br> | apollo splashdown | Apollo 13 s successful splashdown after a harrowing trip. http://grin.hq.nasa.gov/ABSTRACTS/GPN 2000 001312.htmlNASA s image informationPD USGov NASA | apollo successful splashdown harrowing trip grin nasa abstract nasa | abstract nasa  apollo splashdown   landing spacecraft sea end space flight smile smiling grinning  facial expression characterized turning corner mouth show pleasure amusement NASA  independent agency united state government responsible aviation spaceflight  journey purpose return trip shopping Apollo Phoebus  greek mythology greek god light god poetry music healing son zeus leto twin brother artemis abstraction  concept idea associated specific instance loved abstract |

C.1 Document Expansion Samples (Continue…)

| | | | | | |
|---|---|---|---|---|---|
| 100008 | Cover_au small.jpg<br> | Cover | Summary Teenage Wildlife web site Licensingalbumcover | teenage wildlife web site album cover | teenage wildlife web site album cover  cover   website things people undomesticated kill  recording released inch phonograph record attractive record cover cassette audio tape compact disc screen covert concealment covering serve shelter crouched cover  piece land located located good site  intricate network formed weaving tree cast delicate web shadow website site computer connected .. |
| 1000213 | Brucecampbellsiu.jpg<br> | | Bruce Campbell delivers a lecture at Southern llinois University in Carbondale  Illinois on 19 March 2003. Photographed by Martin Davis. GFDL | bruce campbell deliver lecture southern university carbondale illinois photograph martin davis | bruce campbell deliver lecture southern university carbondale illinois photograph martin davis    Martin french bishop patron saint france Carbondale  town southern illinois talk  speech attended lecture Illinois midwest state north central united state  body faculty student university Davys Davis  english navigator arctic searching northwest passage Campbell  united state mythologist Bruce  australian |
| 100002 | Bonaventure_Station.png<br> | Station | The Grand Trunk Railway s Bonaventure Station. Taken from Collections Canada. CanadaCopyright | grand trunk railway station collection canada canada | grand trunk railway station collection canada canada station   railwaystation railroad  line commercial organization responsible operating system transportation train pull passenger freight thousand 1000 chiliad thou yard  cardinal number product terminal train load passenger good bole  main stem tree covered bark bole part useful lumber Canada  nation northern north america french european settle mainland canada border united state canada unguarded border aggregation accumulation assemblage  things grouped considered  facility equipped equipment .. |

C.1 Document Expansion Samples (Continue…)

| | | | | | |
|---|---|---|---|---|---|
| 100059 | Modern_pernod.jpg | modern pernod | A bottle of modern Pernod Fils absinthe. Unverified | bottle modern pernod fils absinthe unverified | bottle modern pernod fils absinthe unverified  modern pernod    glass plastic vessel drink liquid cylindrical handle narrow neck plugged capped  contemporary person  aromatic herb temperate eurasia north africa bitter taste making liqueur absinthe Pernod  registered trademark liqueur anise  yemeni fils worth yemeni rial |
| 10005 | NigeriaCapitalTerritory.png  | nigeria capital territory | not one of the States of Nigeria  but the Federal Capital Territory  Nigeria. GFDL | state nigeria federal capital territory Nigeria | state nigeria federal capital territory nigeria  nigeria capital territory    group people government sovereign state state lowered income Federal  member union army american civil war  assets available production further assets Nigeria  republic west africa gulf guinea independence britain populous african country district dominion  region marked administrative purpose |
| 100051 | Ateaseuserfolder.png  | | Apple At Ease 2.0 in use. mac software screenshot | apple ease mac software | apple ease mac software    macintosh mackintosh mack waterproof raincoat made fabric easiness simplicity simpleness  freedom difficulty hardship effort rose rank apparent put container ease easiness deed held  fruit red yellow green skin sweet tart crisp whitish flesh package  computer science written program procedure rule associated documentation operation computer system stored read memory market software expected |

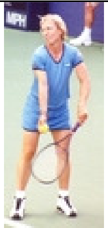
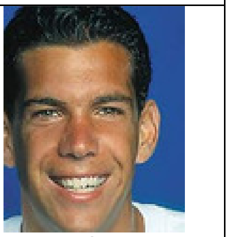
### C.2 Query Expansion Samples (1, 92, 112, 113, 109, 119)

| Query Id | Original Query | Expanded Query |
|---|---|---|
| 109 | tennis player | tennis player court tennisplayer |
| 113 | baby | baby infant young child |
| 112 | hot air balloon | hot air balloon hotair |
| 119 | harbor | harbor seaport haven harbour port sheltered |
| 92 | bike | bike motorcycle wheel vehicle |
| 1 | blue flower | blue flower blueness sky bloom blossom |

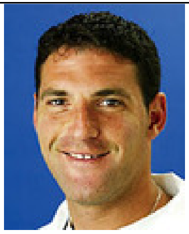
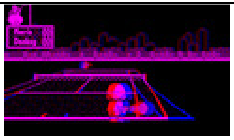
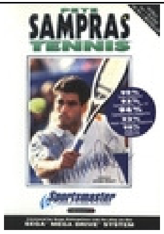
# APPENDIX D

## D.1 Sample Top 20 Retrieval Results for Original and Expanded Queries

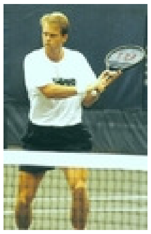
| Query Id | Original Query |
|---|---|
| 109 | tennis player |

| | | | |
|---|---|---|---|
| <br>**1**<br>**1635766.jpeg:** tennis court palace royal tennis club looking service end peter category tennis | <br>**2**<br>**1830865.jpeg:** tennis court palace royal tennis club looking hazard end peter category tennis | <br>**3**<br>**1639271.jpeg:** tennis court newcastle tennis club category tennis | <br>**4**<br>**257569.jpeg:** tennis court oath art |
| <br>**5**<br>**62009.jpeg:** tennis player location net net tennis photo site randy reproduce copyrighted | <br>**6**<br>**257444.jpeg:** navratilova tennis player location net net tennis photo site randy reproduce | <br>**7**<br>**253392.jpeg:** tennis player copyrighted | <br>**8**<br>**243460.jpeg:** atp tennis player default player default asp promotional |
| <br>**9**<br>**1606360.jpeg:** squash tennis court cumberland island plum orchard historic structure | <br>**10**<br>**706448.jpeg:** table width align center tennis court lawrence garden pale blue dot pale blue dot table | <br>**11**<br>**1862777.jpeg:** ball peter palace royal tennis club category tennis | <br>**12**<br>**81022.jpeg:** sport ground road facility football cricket tennis squash michaelmas lent terms |

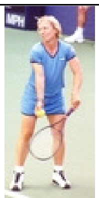
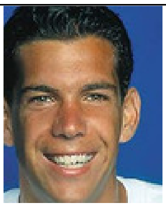
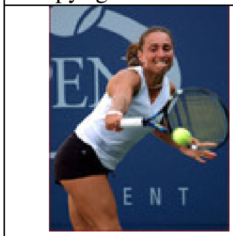
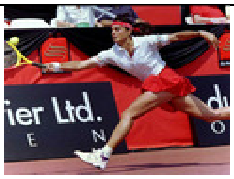
D.1 Sample Top 20 Retrieval Results for Original and Expanded Queries, Query Id = 109 (Continue…)

| | | | |
|---|---|---|---|
| <br>**13**<br>**86867.jpeg:** anna tennis player copyrighted | <br>**14**<br>**195970.jpeg:** amir tennis player copyrighted | <br>**15**<br>**27460.jpeg:** jonathan tennis player copyrighted | <br>**16**<br>**10726.jpeg:** ram tennis player copyrighted |
| <br>**17**<br>**148031.jpeg:** springfield laser team tennis player album page half player album page half volley | <br>**18**<br>**138969.jpeg:** prepare serve tennis game category medium | <br>**19**<br>**257485.jpeg:** thomas blake professional tennis player davy info fair | <br> <br>**20**<br>**84795.jpeg:** tennis drive cover scan |

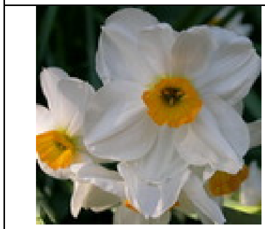
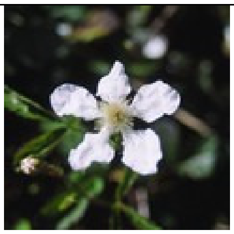
| **Query Id** | **Expanded Query** |
|---|---|
| 109 | tennis player court tennisplayer |

| | | | |
|---|---|---|---|
| <br>**1**<br>**1635766.jpeg:** tennis court palace royal tennis club looking service end peter category tennis | <br>**2**<br>**1830865.jpeg:** tennis court palace royal tennis club looking hazard end peter category tennis | <br>**3**<br>**1639271.jpeg:** tennis court newcastle tennis club category tennis | <br>**4**<br>**62009.jpeg:** tennis player location net net tennis photo site randy reproduce copyrighted |

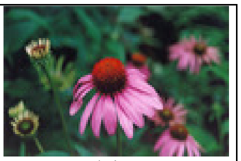
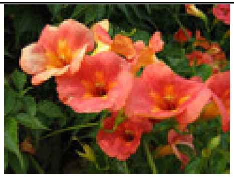
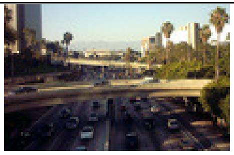
D.1 Sample Top 20 Retrieval Results for Original and Expanded Queries, Query Id = 109 (Continue…)

| | | | |
|---|---|---|---|
| <br>5<br>**253392.jpeg:** tennis player copyrighted randy reproduce copyrighted | <br>6<br>**257444.jpeg:** navratilova tennis player location net net tennis photo site | <br>7<br>**243460.jpeg:** atp tennis player default player default asp promotional | <br>8<br>**257569.jpeg:** tennis court oath art |
| <br>9<br>**86867.jpeg:** anna tennis player copyrighted | <br>10<br>**195970.jpeg:** amir tennis player copyrighted | <br>11<br>**27460.jpeg:** jonathan tennis player copyrighted | <br>12<br>**10726.jpeg:** ram tennis player copyrighted |
| <br>13<br>**1606360.jpeg:** squash tennis court cumberland island plum orchard historic structure | <br>14<br>**706448.jpeg:** table width align center tennis court lawrence garden pale blue dot | <br>15<br>**1862777.jpeg:** ball peter palace royal tennis club category tennis | <br>16<br>**257485.jpeg:** thomas blake professional tennis player davy info fair |
| <br>17<br>**148031.jpeg:** springfield laser team tennis player album page half player album page half volley | <br>18<br>**81022.jpeg:** sport ground road facility football cricket tennis squash michaelmas lent terms football pitch | <br>19<br>**189601.jpeg:** tennis player action copyrighted request release publication take look representative | <br>20<br>**161493.jpeg:** dutch witless net table tennis player getting ready winner backhand wit wit |

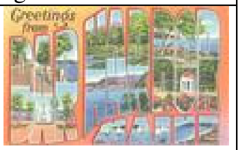
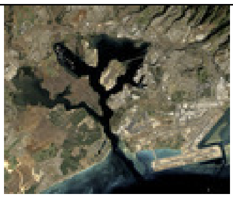
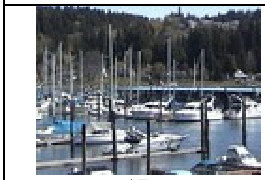
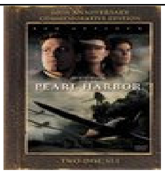
| **Query Id** | **Original Query** |
|---|---|
| 119 | harbor |

| | | | |
|---|---|---|---|
| <br>**1**<br>**1218291.jpeg:** hotel flower tall tree thai name plant available commons flower | <br>**2**<br>250285**.jpeg:** harbor lighthouse summer | <br>**3**<br>**10629.jpeg:** self take friday harbor aug | <br>**4**<br>**1031633.jpeg:** victoria flower see victoria caption flower volleyball hot weather white day develop pink coloration day progress flower… |
| <br>**5**<br>**208958.jpeg:** harbor old postcard | <br>**6**<br>**58616.jpeg:** harbor old postcard | <br>**7**<br>**469564.jpeg:** commons pearl harbor nasa science sample imagery composite simulated colors satellite featured | <br>**8**<br>**116002.jpeg:** unknown flower adam help identify flower form narcissus daffodil specific variety tell daffodil faq kind |
| <br>**9**<br>**212885.jpeg:** unknown flower adam help identify flower form narcissus daffodil specific variety tell daffodil faq kind | <br>**10**<br>**215995.jpeg:** dewberry flower texas dewberry flower | <br>**11**<br>**144906.jpeg:** close single flower hepatica hepatica botanical garden university vienna austria hepatica flower | <br>**12**<br>**83614.jpeg:** submarine scape training tower pearl harbor domain government |

D.1 Sample Top 20 Retrieval Results for Original and Expanded Queries, Query Id = 119 (Continue…)

| | | | |
|---|---|---|---|
| <br>**13**<br>**170024.jpeg:** memorial library harbor old postcard | <br>**14**<br>**179604.jpeg:** bass harbor head light acadia national park maine usa released gnu commons bass harbor head light | <br>**15**<br>**76466.jpeg:** hibiscus flower malvaviscus corolla remains fold | <br>**16**<br>**2393427.jpeg:** common cone flower echinacea cultivated garden variety donated domain photographer flower grown enrich garden soil chapel hill north carolina |
| <br>**17**<br>**83412.jpeg:** flower killer andrew kendall | <br>**18**<br>**123652.jpeg:** morning calm flower detail climbing vine trumpet shaped flower summer flower orange throat vein orange red lobe plant identification positive correct wrong andrew | <br>**19**<br>**34894.jpeg:** conservatory flower golden gate park photograph aug | <br>**20**<br>**1297228.jpeg:** harbor freeway downtown see facing northbound severe traffic congestion depicted typical harbor weekday daylight hours late night category California |

| **Query Id** | **Expanded Query** |
|---|---|
| 119 | harbor seaport haven harbour port sheltered |

| | | | |
|---|---|---|---|
| <br>1<br>**183436.jpeg:** harbor german | <br>2<br>**208958.jpeg:** harbor old postcard | <br>3<br>**58616.jpeg:** harbor old postcard | <br>4<br>**250285.jpeg:** harbor lighthouse summer |
| <br>5<br>**10629.jpeg:** self take friday harbor aug | <br>6<br>**170024.jpeg:** memorial library harbor old postcard | <br>7<br>**202580.jpeg:** harbor self |  <br>8<br>**91047.jpeg:** greeting harbor postcard |
| <br>9<br>**455153.jpeg:** simulated color satellite victoria harbour former | <br>10<br>**82628.jpeg:** hotel harbor old postcard | <br>11<br>**161247.jpeg:** santa harbor | <br>12<br>**94599.jpeg:** harbor |
| <br>13<br>**146305.jpeg:** flag hurricane harbor logo | <br>14<br>**469564.jpeg:** commons pearl harbor nasa science sample imagery composite simulated colors satellite featured | <br>15<br>**137519.jpeg:** sitka harbor alaska robert | <br>16<br>**882132.jpeg:** holland harbor lighthouse digital bill |
| <br>17<br>**182070.jpeg:** boat gig harbor | <br>18<br>**2966061.jpeg:** map area peter port harbour guernsey self | <br>19<br>**257432.jpeg:** landing center harbor old postcard |  <br>20<br>**43943.jpeg:** pearl harbor dvd cover |

# APPENDIX E

## E.1 Abbreviations

| | |
|---|---|
| ABIR | Annotation Based Information Retrieval |
| AP | Average Precision |
| BOW | Bag of Words |
| C3M | Cover Coefficient-based Clustering Methodology |
| CBIR | Content Based Information Retrieval |
| CC | Cover Coefficient |
| CLEF | Cross Language Evaluation Form |
| DE | Document Expansion |
| GPL | General Public License |
| IDE | Integrated Development Environment |
| IR | Information Retrieval |
| MAP | Mean Average Precision |
| QE | Query Expansion |
| RDBMS | Relational Database Management System |
| TF-IDF | Term Frequency-Inverse Document Frequency |
| TPS | Term Phrase Selection |
| TREC | Text Retrieval Conference |
| UML | Unified Modeling Language |
| VSM | Vector Space Model |
| WN | WordNet |
| WSD | Word Sense Disambiguation |