Toward Cross-Language and Cross-Media Image Retrieval*

Carmen Alvarez, Ahmed Id Oumohmed, Max Mignotte, and Jian-Yun Nie

DIRO, University of Montreal, CP. 6128, succursale Centre-ville,
Montreal, Quebec, H3C 3J7 Canada
{bissettc, idoumoha, mignotte, nie}@iro.umontreal.ca

Abstract. This paper describes our approach used in ImageCLEF 2004. Our focus is on image retrieval using text, i.e. Cross-Media IR. To do this, we first determine the strong relationships between keywords and types of visual features such as texture or shape. Then, the subset of images retrieved by text retrieval are used as examples to match other images according to the most important types of features of the query words.

1 Introduction

There has been a large amount of research on Cross-Language Information Retrieval (CLIR), but much less for Cross-Media Information Retrieval (CMIR). By CMIR, we mean that a user issues a query in one medium (e.g. in natural language), but retrieves information in another medium (e.g. images). CMIR is similar to CLIR. The key problem is still (query) translation, but from one medium to another, instead of from one language to another. We call this problem media translation (in contrast to language translation). Media translation is more difficult than language translation. For language translation, there are a large number of resources and tools. In each language, for the purposes of IR, one can reasonably assume that text can be decomposed into words, and there is a finite number of words in each language. There is also a great similarity in the meanings that one can express in different languages, despite the fact that some languages can express meanings that cannot be expressed in others. This means that translation between languages is a feasible, although difficult, task.

For CMIR, the picture is different. There is no equivalent of words in languages for images. Indeed, we do not have a finite number of features that are semantically meaningful and which can be used to characterise the semantics of each image. Representing an image by a set of pixels does not help to understand what the image "means". In such a situation, it is impossible to build a system equivalent to Machine Translation (MT) between natural languages. It is impossible to even build a kind of "bi-media" dictionary that maps a word to

* This study has been partly supported by the Center of Interdisciplinary Research en Emerging Technologies (CITÉ) of the University of Montreal.

C. Peters et al. (Eds.): CLEF 2004, LNCS 3491, pp. 676–687, 2005.

one or several image components or characteristics. How can one build a CMIR system that automatically "translates" a text query into an image query? This is the key problem that we investigate in our study. Our approach is inspired by CLIR approaches based on parallel corpora. Using a parallel corpus, one can train a statistical translation model between two languages. We notice that it is relatively easy to obtain such a parallel corpus between text and images: any set of images annotated with words, or accompanied by text, can be regarded as such a parallel corpus between images and text. The problem is to have a reasonable approach to train a translation model.

Our initial goal in this research was to build a translation model as in natural languages. We intended to extract image characteristics that can play a similar role to words in text and then, train a statistical translation model between words and these characteristics. However, this task proved too ambitious. The approach used in our experiments for ImageCLEF 2004 was to use the parallel texts in a manner similar to that of Yang et al. [1]. The original query is used to retrieve a set of documents from the parallel corpus in the source language. Target words are extracted from parallel texts of these documents, and the target words are used as query translation. We also use a text query (one keyword) to retrieve the images that contain the keyword in annotation. Then, the common characteristics of these images (their centroid) are used as image query. The results of this approach seem promising: although the effectiveness of the CMIR approach alone is far lower than using text queries to match text annotations, we observe that the images retrieved for keywords with strong visual characteristics such as "garden" and "boat" are relevant even if not annotated with these words. Such a CMIR approach allows us to extend retrieval results to cover the relevant images without text annotation, which is the case for most images on the Web.

2 Existing Image Retrieval Approaches

Different approaches have been used for image retrieval. A user can submit a text query and the system can search for images using image captions. A user can submit an image query (using an example image - either selected from a database or drawn by the user). In this case, the system tries to determine the most similar images to the example by comparing various visual features such as shape, texture, or color. A third group of approaches tries to assign semantic meaning to images. This approach is often used to annotate images by concepts or keywords [2]. Once images have been associated with keywords, they can be retrieved by text. The approaches have their own advantages and weaknesses. The first approach is indeed text retrieval. There is no image processing. Coverage of the retrieval is limited to images with captions or annotation, which can be explicit or implicit (file name). The second approach does not require images to be associated with captions. However, the user is required to provide an example image and a visual feature or a combination of some features to be used for comparison. This is often difficult for a non-expert user and the retrieval effectiveness (for high-level queries) is lower than for text retrieval. The third approach, if successful, allows us to automatically recognise the semantics of images, and thus allow users to query images by keywords. Currently, only annotation of images by to typical components or features seems possible. For example, according to a texture analysis, one can recognise a region of images as corresponding to a tiger due to the texture of tigers [3]. The particular feature to be used is selected manually according to the objects to be annotated. It is still impossible to recognise all semantic meanings of images.

Recent studies [4] have tried to automatically create associations between visual features and keywords. The basic idea is to use a set of annotated images as a set of learning examples, and to extract associations between annotations and the visual features of images. In our study, we initially tried to use a similar approach in ImageCLEF. That is, we wanted to extract strong relationships between the keywords in the captions and the visual features of the images. If such relationships could be created, then it would be possible to use them to retrieve non-annotated images by a textual query. In this case, the relationships play a role of translation between media. However, we discovered that this approach is extremely difficult in the context of ImageCLEF for several reasons:

1. The annotations in the ImageCLEF corpus often contain keywords that are not strongly associated with particular visual features. They correspond to abstract concepts. Examples of such keywords are "Scotland", "north", and "tournament". If we use the approach systematically for every word, there will be noisy relationships.
2. Even if there are relationships between keywords and visual features, the relationships may be difficult to extract because there are a huge number of possible visual features. In fact, visual features are continuous. Even if we use some discretisation techniques, their number is still too high to be associated with keywords. For example, for a set of images associated with the keyword "water", one would expect to extract strong relationships between the keyword and the color and texture features. However, "water" in images may only take up a small region of the image. There may be other objects, making it difficult to automatically isolate the typical features for "water".

Due to this, we take a more flexible approach. We also use images with captions as a set of training examples, but we do not try to create relationships between keywords and particular visual features (such as a particular shade of blue for the word "water"). We only try to determine which type(s) of feature are the most important for a keyword. For example, "water" may be associated with "texture" and "color". Only strong relationships are retained. During the retrieval process, a text query is first matched with a set of images using captions. This is a text retrieval step. Then, the retrieved images are used as examples to retrieve other images, which are similar according to the determined types of features associated with the keywords. The process of our system is illustrated in Figure 1.

This approach, if successful, is very useful in practice. In many cases, image captions contain abstract keywords that cannot be strongly associated with visual features, and even if they can, it is impossible to associate a single vector to a keyword. Our approach does not require determining such a single feature

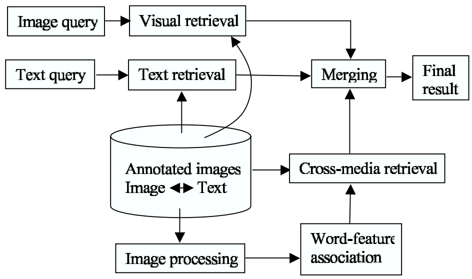

Fig. 1. Workflow of image retrieval

vector for a given keyword. It abandons the third approach mentioned earlier, but combines the first two approaches. The advantage of extracting keyword-feature associations is to avoid the burden of requiring the user to indicate the appropriate types of features to be used in image comparison.

3 Image Processing-Based Learning Procedure

Objective of the automatic image processing-based learning procedure is:

- to determine the most discriminant type(s) of high-level visual features for each annotated keyword. We have considered the three fundamental visual characteristics; namely, *texture* (including color information), *edge* and *shape*. For example, the keyword "animal" could belong to the *shape* class since the measure using *shape* information will be the most discriminant to identify images with animals (although "zebra" and "tiger" will probably belong to the *edge* and *texture* classes).
- to identify candidate images that are most representative for a keyword.

The type of high-level visual feature along with its discriminant measure and a set of representative images will be used to refine the image retrieval process.

3.1 Edge Class and Its Measure

Wavelet-based measures have often been used in content-based image retrieval system because of the appealing ability to describe local texture and the distribution of the edges of a given image at multiple scales. We use the Harr wavelet transform [5] for the luminance (grey-level) component of the image. This transform is chosen for its better localisation properties and it requires less computation compared to others (e.g., Daubechies' wavelet). The image decomposition into wavelets involves recursive numeric filtering. At each step in the recursion, we obtain four sub-bands, which we refer to as LL, LH, HL and HH according to their frequency characteristics (L: Low and H: High). The LL sub-band is then decomposed into four sub-bands at the next decomposition step. We use three steps in our case. For each sub-band, we compute the mean and the standard deviation of the energy distribution. This leads to a vector of 20 (i.e., $(2 \times 3 \times 3) + 2$) components considered as the descriptors (or signature) of the *edge* characteristics.

3.2 Texture Class and Its Measure

Tamura *et al.* [6] proposed to characterise image texture along the dimensions of contrast, directionality, coarseness, line-likeness, regularity and roughness. In this class of visual features, we use only the coarseness property which yields a histogram of six bins, for the following reasons:

- Contrast is not very discriminant for texture retrieval,
- Edge information is already treated in the wavelet and shape class,
- Line-likeness, regularity and roughness are correlated to coarseness, contrast and directionality.

Coarseness refers to the size of the *texton*; the smallest unit of a texture. This measure depends on the resolution of the texture. With this measure, we can compute a histogram with 6 bins (a 6-component attribute vector), which will be used as the descriptor of the *texture* of a given image.

3.3 Shape Class and Its Measure

Description and interpretation of shapes in an input image remains difficult. Several methods use a contour detection in the images (such as Canny or Sobel edge detectors) as a preliminary step. These methods remain dependent on parameters such as thresholds (on the magnitude of the image gradient). In image compression, a vector quantisation method is used [7] on the set of vectors of dimension K^2 of grey-levels corresponding to $K \times K$ blocks extracted from the image. By using a clustering procedure into K classes, we can obtain an image with separate regions (a set of connected pixels belonging to the same class) from which we extract the contours of regions. These edges are connected and obtained without parameter adjustment and noise is taken into consideration. We use this strategy of edge detection and as clustering procedure, we use the Generalised Lloyd [8] [9] method. In our implementation, we use the QccPack Library[1]. For each edge pixel, we define a direction (horizontal, vertical, first or second diagonal) depending on the disposition of its neighbouring edge pixels. For each direction we count the number of edge pixels associated with it, which yields a 4 bin histogram.

4 Relationships Between Words and Visual Features

4.1 The Learning Procedure

The learning procedure determines the type of high-level visual features that are most representative for each annotated keyword:

[1] `http://qccpack.sourceforge.net/`

Fig. 2. *Result of learning procedure applied to "garden". Below each image, we put the ID of the image and its similarity score. If the image is not annotated by the word "garden", its ID is put in a gray box.*

1. Let $\mathbf{I}_w$ be the set of all images I_w (each described by its three vectors or descriptors $[D_{I_w}^{texture}, D_{I_w}^{edge}, D_{I_w}^{Shape}]$) in the training database that are annotated with the keyword w and $|\mathbf{I}_w|$, the number of images in $\mathbf{I}_w$.
2. For each CLASS { *Texture*, *Edge*, *Shape* }
   (a) We use a K-mean clustering procedure [10] (euclidean distance as similarity measure) on the set of samples $D_{I_w}^{class}$.
   (b) This clustering allows us to approximate the distribution of the set of samples $D_{I_w}^{class}$ by K spherical distributions (with identical radius) and to give K prototype vectors $[D_{1,w}^{class}, ..., D_{k,w}^{class}]$ corresponding to the centers of these distributions. Several values of K are used to find the best clustering representation of $D_{I_w}^{class}$.
   (c) For each PROTOTYPE VECTOR $\{ D_{1,w}^{class}, ..., D_{k,w}^{class} \}$
      - We search in the training database for the closest descriptors (or images) of $D_{k,w}^{class}$, according to euclidean distance. Let $\mathbf{I}_{k,w}^{class}$ be this set of images.
      - We compute the number of the first top-level T samples of $\mathbf{I}_{k,w}^{class}$ also belonging to $\mathbf{I}_w$ (best results obtained with $T = 10$). Let $N_{k,w}^{class}$ be this number.
3. We retain the CLASS(ES) and $\mathbf{I}_{k,w}^{class}$ for which we have $N_{k,w}^{class}$ above a given threshold ξ. At this point, a keyword may be associated with some strong class(es) of features as well as a typical image cluster (i.e. $\mathbf{I}_{k,w}^{class}$).

4.2 Cross-Media Retrieval

The simplest case of CMIR is with only one keyword. For image retrieval with a keyword, the set of the associated image clusters $\mathbf{I}_{k,w}^{class}$ is used as example image to match the whole image database, according to the feature class(es). For a text query with several keywords, this process is repeated for each keyword. The lists of retrieved documents (images) are first normalized by relevance score (i.e. the relevance score is divided by the maximum score in the list) and then merged with equal importance (1/3 for each list). The final list is the result of CMIR.

We observe that this method works well for keywords with strong visual features such as "garden" and "boat". The first 24 images of the set of images associated to the word *garden* are shown in Figure 2. We can see that, even if most images are not annotated by the word *garden* (the word does not exist in any field of the text associated with the image), we can visually count about 9 images related to gardens from the 14 non-annotated images.

5 Cross-Language Text Retrieval

5.1 Translation Models

Two approaches are used for query translation, depending on the resources available for the different languages. For Spanish, Italian, German, Dutch, Swedish, and Finnish, FreeLang bilingual dictionaries [2] are used in a word-for-word translation approach. The foreign language words in the dictionaries are stemmed using Porter stemmers[3], and the English words are left in their original form. The queries are also stemmed, and stop words are removed with a stoplist in the language. The translated query consists of the set of all possible English word translations for each query term, each translated word having equal weight. For French, a translation model trained on a web-aligned corpus is used [11]. The model associates a list of English words and their corresponding probabilities with a French word. As with the bilingual dictionaries, the French words are stemmed, and the English words are not. Word-for-word translation is done. For a given French root, all possible English translations are added to the translated query. The translation probabilities determine the weight of the word in the translated query. The term weights are represented implicitly by repeating a given translated word a number of times according to its translation probability. For French as well as for the other languages, the words in the translated query are stemmed using the Porter stemming algorithm.

5.2 CLIR Process

For retrieval, the Okapi retrieval algorithm [12] is used, implemented by the Lemur Toolkit for Language Modeling and Information Retrieval. The BM 25 weighting function is used. We also use pseudo-relevance feedback. The parameters are trained on the TREC-6 AP89 document collection and 53 queries in English, French, Spanish, German, Italian, and Dutch. Since no training data was available for Finnish and Swedish, the average of the optimal values found

[2] `http://www.freelang.net`

[3] `http://snowball.tartarus.org`

for the other languages is used. While the training collection, consisting of news articles about 200-400 words in length, is quite different from the test collection of image captions, the volume of the training data (163'000 documents, 25 or 53 queries, depending on language, 9403 relevance assessments) is much greater than the training data provided by the image collection (5 queries, 167 relevance assessments). For our experiments, the above parameters are set as follows:

- Okapi parameters: k1=1.0-1.5, b=0.75-0.8 and k3=7-9;
- Feedback parameters: FeedbackDocCount=20, FeedbackTermCount=5 and qtf (weight of query terms added) =0.2-0.6.

Given a text query, we retrieve a list of images using the above parameters. We annotate this image relevance score based on textual retrieval as $R_{text}(i, q)$.

6 Combining Text and Images in Image Retrieval

6.1 The Image Relevance Score Based on Clustering

The image analysis based on clustering, described in section 4, provides a list of retrieved images i for a given word w, with a relevance score for each image, $R_{cluster}(i, w)$. The relevance score, based on clustering, is a weighted sum of the relevance scores for that image for each (non stopword) query term:

$$R_{cluster}(i, q) = \sum_{w \in q} \lambda_w R_{cluster}(i, w) \tag{1}$$

Each word has the same weight and the relevance score for the query is normalised with $\lambda_w = \frac{1}{|q|}$, where $|q|$ is the number of words in the query.

6.2 Image Retrieval Using Image Queries

In ImageCLEF, we are also provided with one example image for each topic. These images can be used for content-based image retrieval using visual features. We use the same visual features (i.e. edge, texture and shape) as described in 3 for the calculation of image similarity. Using the three classes of visual features, the following relevance scores are obtained: $R_{edge}(i, q)$, $R_{texture}(i, q)$ and $R_{shape}(i, q)$. They are merged into a visual similarity score as follows:

$$R_{visual}(i, q) = \lambda_{edge} R_{edge}(i, q) + \lambda_{texture} R_{texture}(i, q) + \lambda_{shape} R_{shape}(i, q) \tag{2}$$

In our experiments, we give equal importance to the three visual features.

6.3 Combining the Five Image Relevance Scores

We now have 5 lists of images, with the following three types of scores:

- $R_{text}(i, q)$: Text retrieval score;
- $R_{cluster}(i, q)$: Cross-media retrieval score;
- $R_{edge}(i, q)$, $R_{texture}(i, q)$, $R_{shape}(i, q)$: Visual similarity scores.

Each of these relevance scores contributes to the final relevance score as follows:

$$R(i,q) = \lambda_{text} R_{text}(i,q) + \lambda_{cluster} R_{cluster}(i,q) + \lambda_{visual} R_{visual}(i,q) \quad (3)$$

The coefficients chosen for the contribution of each approach are as follows: $\lambda_{text} = 0.8$, $\lambda_{cluster} = 0.1$, $\lambda_{visual} = 0.1$. These values have been determined empirically using the training data.

6.4 Filtering Images Based on Location, Photographer, and Date

A final filtering is applied to the list of images for a given query for location, photographer, and date, when these latter are specified in the query. These entities were extracted from the data associated with the images.

7 Experiments

7.1 Monolingual and Bilingual Text Retrieval

Our experiments only use topic titles. For query translation, English-French translation is performed with a statistical translation model trained on a set of parallel Web pages. For other languages, the translation is done with bilingual dictionaries. Table 1 shows the effectiveness obtained for monolingual text retrieval (E-E) and bilingual retrieval with French (F-E) and Spanish (S-E) queries.

Table 1. Effectiveness of text retrieval

	F-E	E-E	S-E
Title	0.4976	0.5530	0.4843

The above results were obtained with filtering by date, place and photographer when this is specified in the query. Without filtering, we observe a decrease for F-E and S-E experiments: 0.4838 for F-E and 0.4513 for S-E. For monolingual retrieval (E-E), without filtering, the effectiveness is slightly higher: 0.5729.

7.2 Cross-Media Retrieval

If we only use the CMIR method (or the method based o clustering) we developed, we obtain the effectiveness shown in table 2:

Notice that in these experiments, the queries are still written in different languages. So they have to be translated into English before the CMIR method is used. The effectiveness for F-E and S-E is indeed a combined effectiveness of CLIR and CMIR. We can observe that the effectiveness obtained is much lower than with text retrieval, although visual retrieval performs better than CMIR. This clearly shows that the CMIR method cannot be used alone for image retrieval.

Table 2. Effectiveness of CMIR

E-E	F-E	S-E
0.0536	0.0486	0.0321

Table 3. Effectiveness of combined approaches

T(0.8)+C(0.2)	T(0.8)+V(0.2)	T(0.8)+C(0.1)+V(0.1)		
E-E	E-E	E-E	F-E	S-E
0.5502	0.5699	0.5620	0.5125	0.4890

7.3 Visual Retrieval

If we use the example images provided with the queries to retrieve similar images, we obtain an average precision of 0.0586. If we use filtering by date, place and photographer, we have to use CLIR to some degree. Then, the effectiveness for E-E, F-E and S-E is respectively 0.0999, 0.0952 and 0.0972. The difference in language translation has almost no impact on the filtering process.

7.4 Combined Approaches

As both CMIR and visual retrieval have low effectiveness and cannot be used alone for image retrieval, we combine these methods with text retrieval in order to improve the latter. The following table shows the combinations and the effectiveness obtained (where T(X)+C(Y)+V(Z) means that text retrieval, CMIR and visual retrieval are signed an importance X, Y and Z):

We can see in the third column that when CMIR and visual retrieval are combined with text retrieval, the latter can be further improved. We tested with different importance values for the three types of retrieval. It turns out that text retrieval should be attributed with a high importance value (above 0.6) for the combined approach to be effective. The two other retrieval methods should be given low importance (about 0.1). These importance values are consistent with the effectiveness level of each retrieval method.

In order to see the impact of visual retrieval and CMIR, we combine these two methods separately with text retrieval (first two columns) for English queries. It turns out that when visual retrieval is combined with text retrieval to some extent, the effectiveness can be slightly improved (from 0.5530 to 0.5699). The combination of 0.8 for text retrieval and 0.2 for visual retrieval seems to be the best one. On the other hand, when CMIR is combined with text retrieval, the effectiveness seems to decrease (from 0.5530 to 0.5502). This result may suggest that our current way to do CMIR for querying is inappropriate. Indeed, for a query, we currently retrieve a list of images for each single word, then combine them with equal importance. It may be better to consider the relative importance of each word in the query. Our examination on some of the lists retrieved for keywords such as "garden" suggests that our CMIR method may work well for these words with strong relationships with visual features. On the other hand, more abstract words such as "tournament" are not connected with any particular visual features. It would be appropriate to use CMIR for the first group of words, but not for the second group. We will investigate the proper utilisation of our CMIR approach in the future.

8 Discussions

In this study, we propose a method to automatically extract relationships between keywords and visual features. The extraction approach is inspired by the CLIR approach based on parallel corpora. For image retrieval with words that are strongly related to visual features, this method seems to work well. However, when CMIR is used in a simplistic way for a query with several words in our ImageCLEF experiments, the method does not seem to bring any positive impact. We also tested the combination of image retrieval with both text and image queries. Such a combination brings improvements in comparison with text retrieval alone. The current implementation is still quite simple, and the idea of CMIR using annotated images as a parallel corpus is not yet fully tested. There are several possible improvements that we can do in the future:

- In our current experiments, we retrieve a list of documents (images) for each keyword in a text query, and we assign equal importance to all the keywords. In fact, it would be possible to attribute a higher importance to a keyword that is judged more important, or related more strongly to some particular visual features. In this way, we will be able to rely more on keywords such as "garden" and "water", and less on "golf" and "tournament".
- In our visual retrieval, we attributed the three classes of features an equal importance. It would be possible to assign different importance according to the keywords in the query. For example, if the keywords are more related to texture than to shape, then the texture similarity could be given higher importance in the merged result.
- In our experiments on CLIR with French queries, it turns out that using short queries (titles) are better than using long queries (titles and narratives). It is related to the number of translation words that we select for the translation. When we translate with a statistical translation model, the number of translation words to be selected is important. The higher effectiveness with titles suggests that the translation of query with a translation model indeed produce a desirable query expansion effect - the effect that has been mentioned in several previous studies on CLIR. As a consequence, statistical translation models could be particularly adapted to short queries.

In our future research, we will further investigate the CMIR approach in order to understand how words should be translated into visual features.

References

1. Yang, Y., Carbonell, J., Brown, R., Frederking, R.: Translingual information retrieval: learning from bilingual corpora. Artificial Intelligence **103** (1998) 323–345

2. Liu, W., Dumais, S., Sun, Y., Zhang, H., Czerwinski, M., Field, B.: Semi-automatic image annotation (2001)
3. Carson, C., Thomas, M., Belongie, S., Hellerstein, J.M., Malik, J.: Blobworld: A system for region-based image indexing and retrieval. In: Third International Conference on Visual Information Systems, Springer (1999)
4. Jeon, J., Lavrenko, V., Manmatha, R.: Automatic image annotation and retrieval using cross-media relevance models. In: ACM SIGIR. (2003)
5. Mallat, S.: A theory for multiresolution signal decomposition : The wavelet representation. IEEE Transactions on Pattern Analysis and Machine Intelligence **11** (1989) 674–693
6. Tamura, H., Mori, S., , Yamawaki, T.: Texture features corresponding to visual perception. IEEE Transactions on Systems, Man, and Cybernetics **8** (1978) 460–473
7. Goldberg, M., Boucher, P., Shlien, S.: Image compression using adaptative vector quantization. Communications, IEEE Transactions on [legacy, pre - 1988] **34** (1986) 180–187
8. Lloyd, S.: Last square quantization in pcm's. Bell Telephone Laboratories Paper (1957)
9. Linde, Y., Buzo, A., Gray, R.: An algorithm for vector quantizer design. IEEE Transactions on Communications **COM-28** (1980) 84–95
10. Banks, S.: Signal processing image processing and pattern recognition. Prentice Hall (1990)
11. Nie, J.Y., Simard, M.: Using statistical translation models for bilingual ir. In: Revised Papers from the Second Workshop of the Cross-Language Evaluation Forum on Evaluation of Cross-Language Information Retrieval Systems. (2001) 137–150
12. Robertson, S., Walker, S., Jones, S., Hancock-Beaulieu, M., Gatford, M.: Okapi at trec-3. In: Proc. of the Third Text REtrieval Conference (TREC-3), NIST Special Publication 500-225. (1995)