COMBINING TEXTUAL AND VISUAL ONTOLOGIES TO SOLVE MEDICAL MULTIMODAL QUERIES

Saïd Radhouani[1], Joo Hwee Lim[2], Jean-Pierre Chevallet[2], Gilles Falquet[1]

[1]Centre Universitaire d'Informatique, 24, rue Général-Dufour, CH-1211 Genève 4, Switzerland
[2]IPAL Lab, Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613
Said.Radhouani@cui.unige.ch

ABSTRACT

In order to solve medical multimodal queries, we propose to split the queries in different dimensions using ontology. We extract both textual and visual terms depending on the ontology dimension they belong to. Based on these terms, we build different sub queries each corresponds to one query dimension. Then we use Boolean expressions on these sub queries to filter the entire document collection. The filtered document set is ranked using the techniques in Vector Space Model. We also combine the ranked lists generated using both text and image indexes to further improve the retrieval performance. We have achieved the best overall performance for the Medical Image Retrieval Task in CLEF 2005. These experimental results show that while most queries are better handled by the text query processing as most semantic information are contained in the medical text cases, both textual and visual ontology dimensions are complementary in improving the results during media fusion.

1. INTRODUCTION

In specific domains such as medicine, the task of information retrieval is rather specific. Indeed, in this kind of domain, the vocabulary is precise and less ambiguous. When one wants to seek information, he can express his information need through a precise query. Such queries require specific processing to obtain precise document answer. In this paper, we investigate which effects can be achieved for information retrieval by integrating explicit knowledge although an ontology to process precise queries. We add semantics from ontology to handle a rich query language consisting of using query dimensions and Boolean operators. We also present how the fusion of ranked lists, generated using both text and image indexes, can contribute to solve precise multimodal queries and improve the retrieval performance.

In the rest of this paper, we first present our problem through an example of a multimodal query, and then some related works. We introduce the text query module, and the image query module, respectively in sections 3 and 4. We present in section 5 the technique used to fuse text and images. For the evaluation (cf. section 6), we have investigated the medical CLEF-2005 collection. Finally, we conclude and discuss our future works (cf. section 7).

2. MEDICAL MULTIMODAL QUERIES

The example shown in Figure 1 is one of the 25 queries of the CLEF Medical Image Retrieval Task [5]. The test collection used in this task contains images with annotations in XML format. Each query is composed of example query images and text descriptions in natural language.

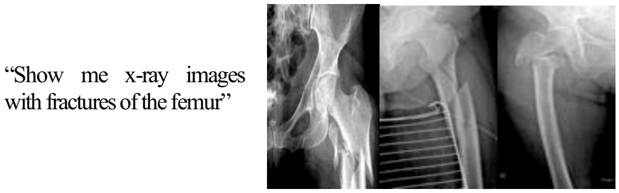

Fig. 1. A typical multi-modal query in the Medical Image Retrieval Task

In this query, it is clear for a human reader that we are looking for images that contain two elements: one part of the **anatomy**, namely a *femur*, and one **pathology**, namely *fracture*. These two elements are semantically related. The fracture is a pathology of a bone such as the femur. These two elements should be described in images whose **modality** is *x-ray*. Thus, images that contain "a fracture of a cranium", or "a femur without fracture" are not relevant to this query. By observing the set of queries in CLEF collection, we noticed that almost queries have these elements (anatomy, pathology, and modality). Hence we call these elements the **dimensions** of the query, and we make the assumption that a relevant document, to one query with dimensions, is the one that fulfils correctly to these dimensions.

In order to take into account this concept of dimensions, initially it is necessary to define and identify them. The dimensions depend on the organization of the studied domain and can be described through external resources (thesaurus, ontology, etc.). These resources contain terms that describe the dimensions. Thus, in order to identify dimensions from queries/documents, we need to extract all terms from each query/document and verify to which dimension does each concept belong.

As one document can contain more than one instance of the same dimension, the presence of all the query dimensions in a document does not imply systematically that such document is relevant. For this reason, we need to take into account the relation between the query dimensions at the document level. This can be done by taking advantage of the rich information expressed in the example query images. Indeed, images contain all query dimensions (*Anatomy*, *Pathology*, and *Modality*), and also relations among them that cannot be fully transcribed into text.

We also notice that there are many things that are invisible to the untrained eye or that can not be noticed without knowing it. For example, slight fractures in bones can be very hard to see on an x-ray image as they are often very thin. Thus, we think that textual information is inseparable from image data. Also, a large fraction of medical image data becomes meaningless without its associated textual descriptions. For all these reasons, we propose to take into account both text and images to improve the retrieval results.

Thus, to resolve medical multimodal queries, it is clear that at first, we have to process each textual query in order to extract its **dimensions** and take them into account into an IR system. Then, we have to process the query image field to get complementary information to query text field.

Several works have proposed approaches to combine text with images in the image retrieval task. In [11], authors incorporate image annotations, thus, the combination is done by query expansion: the first query is done image-

1-4244-0367-7/06/$20.00 ©2006 IEEE

based only. Then, the first k results are used for text-based retrieval. The resulting scores are weighted and combined to the final score.

Jeon et al. propose a cross-media relevance model to learn joint distribution of blobs and words. They further proposed continuous-space relevance model that learning the joint probability of words and regions, rather than blobs [7].

In [10], authors define a meta-language for representing text and images. To transform images to this language, they segment all images into regions and cluster the set of all regions into blobs. Then, they create a co-occurrence model of blobs and words in the textual image descriptions. The indices of the blobs are fed into an information retrieval engine. Now, image- and text-information can be used for retrieval.

Alvarez et al. calculate feature vectors for certain properties (texture, shape and edge) of each image [1]. These features are used for image retrieval, together with a model that assigns each description term a kind of feature that it may refer to. These correlations are incorporated in the retrieval process.

In the next sections, we present how we combine text and images to solve the multimodal queries.

3. TEXT QUERY PROCESSING

In this paper, we propose to use of the Vector Space Model (VSM) to index textual documents/queries. There is no way for this model to take into account the notion of query dimensions because the queries are considered as bags of words [12]. One possibility is to add semantics to the terms query before given them as input of the VSM. Such semantics can be added from external knowledge given through ontology.

The dimensions of one domain are relative to the ontology: a dimension O_{di} is a sub tree of ontology. Thus, a dimension contains all terms belonging to the corresponding sub tree.

Several works proposed to use ontology during the querying process, in particular, they expand the original query by knowledge given through the ontology [2,13,14]. In next section, we propose our ontology-based approach for solving multidimensional queries.

3.1. Using ontology to take into account the query dimensions

Let us represent a query $Q := (Q_{TXT}, Q_{IMG})$, where Q_{TXT} is the text field, and Q_{IMG} is the image field. $Q_{TXT} := \{t_1, \ldots, t_n\}$ is the set of all different terms occurring in the text field. Extracting the query dimensions consists to split the text field Q_{TXT} into different sub queries, each one corresponds to one dimension *di*. Thus, we do a mapping between the text field Q_{TXT} and each ontology dimension O_{di}. Hence, we obtain, for each ontology dimension O_{di}, a query dimension $Q_{di} := \{t_{di}\}$, where t_{di} is a term occurring in the query text field and in the ontology dimension O_{di}. We do not solve yet any term ambiguity because the query belongs to a precise domain. Finally, the text field is represented as follow $Q_{TXT} := \{Q_{d1}, \ldots Q_{di}, \ldots Q_{dn}\}$, where Q_{di} is the sub query corresponding to the dimensions *di*, and n is the number of dimensions occurring in the query.

Once the query dimensions extracted, we do a filtering to determine documents that contain them. Thus, we query the whole document collection using Q_{di} and select document in which at least one term of Q_{di} appears and obtain a sub set D_{di} of the collection. In order to solve the original multidimensional query, we finally combine its dimensions using a Boolean expression. A conjunction forces dimensions to be present together in the document. We can reduce this constraint using a disjunction. We compute this Boolean dimension constraint formula using all sub sets $\{D_{di}\}$. For example, for an initial query text field Q_{TXT} containing three dimensions, sub queries Q_{d1}, Q_{d2} and Q_{d3} are build, and D_{d1}, D_{d2}, D_{d3} are obtained. If we decide that a relevant document must include dimension *d1* and dimension *d2* or only dimension *d3*, we compute the sub document set by the Boolean formula $D_{TXT} = (D_{d1} \cap D_{d2}) \cup D_{d3}$.

After this filtering, the next step is to rank the obtained sub set D_{TXT} with respect to the query text field Q_{TXT}. Thus, we query it using the full original query text field Q_{TXT} using the VSM. Finally, we have the sub set $\mathrm{D}_{\mathrm{TXT}}$ that represents a ranked list of document obtained through the process of the query text field. This method is an extension of the approach we proposed in multilingual CLEF 2005 [6].

It is clear that the query text field is precise specifying explicitly the dimensions, but it does not contain enough information to describe the relations between them. Thus, we consider that query image field contains complementary information to the text field, and we propose to use it during the retrieval process. Our principal hypothesis is that documents retrieved both by text processing and image processing are most relevant than document retrieved only by one separate media. Hence, we propose to use the query image field Q_{IMG}. Our goal is also to show that fusion of image and text can give better results than separate results. In the following section we describe our approach of image processing.

4. IMAGE QUERY PROCESSING

We have applied the VisMed approach [8] on the CLEF Medical Image Retrieval task. We designed 39 VisMed terms that are both relevant to the 25 query topics and that correspond to typical semantic regions in the medical images. Fig. 2 illustrates one visual example each from these 39 VisMed terms.

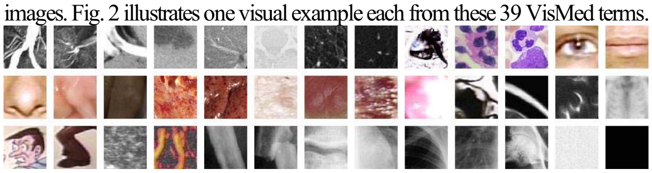

Fig. 2. Visual examples (one each) for the 39 VisMed terms

Based on 0.3% (i.e. 158 images) of the 50,026 images from the 4 collections plus 96 images obtained from the web (i.e. to minimize the number of images selected from the test collection), we manually cropped 1460 image regions to train and validate the VisMed terms using SVMs. For a given VisMed term, the negative samples are the union of the positive samples of all the other 38 VisMed terms. We ensure that they do not contain any of the positive and negative query images given by the query topics. The odd and even entries of the cropped regions are used as training and validation sets respectively (i.e. 730 each) to optimize the RBF kernel parameter of support vector machines. Both the training and validation sets are then combined to form a larger training set to retrain the 39 VisMed detectors.

During image indexing, the 39 trained VisMed detectors are applied to the small regions in a medical image obtained in a sliding window manner [8,9]. After multi-scale reconciliation [8,9], an image block Z in a grid-based image index contains a vector of detection confidence values of VisMed terms, $T_i(Z)$, which is aggregated (averaged) over the detection vector on smaller regions z_k,

$$T_i(Z) = \frac{1}{n}\sum_k T_i(z_k) \tag{1}$$

4.1. Similarity-Based Retrieval with Visual Query

Given two images represented as different grid patterns of $T_i(Z)$, we developed a flexible tiling (FlexiTile) matching scheme [8] to cover all possible matches. For instance, given a query image Q_{IMG} of 3 X 1grid and an image Z of 3 X 3 grid, intuitively Q_{IMG} should be compared to each of the 3 columns in Z and the highest similarity will be treated as the final matching score. In this paper, we denote the similarity between query images Q_{IMG} and database image Z as $\lambda(Q_{IMG}, Z)$.

4.2. Semantics-Based Retrieval with Text Query

A new visual query language, Query by Spatial Icons (QBSI), has been developed to combine pattern matching and logical inference [9]. In this paper, we extend QBSI with spatial quantifiers and apply it to medical semantics-based retrieval where the queries are text description and the query processing is carried on image indexes based on VisMed terms.

A QBSI query is composed as a spatial arrangement of visual semantics. A Visual Query Term (VQT) *P* specifies a region *R* where a VisMed *i* should appear and a query formulas chains these terms up via logical operators. The truth value μ(*P*,*Z*) of a VQT *P* for any image *Z* is simply defined as

$$\mu\left(P,\mathrm{Z}\right)=T_i(R) \tag{2}$$

where $T_i(R)$ is defined in Eq. (1).

In our experiments, the medical images are indexed as 3 X 1, 3 X 2, 3 X 3, 2 X 3, and 1 X 3 grids, depending on their original aspect ratios. When a query involves the presence of a VisMed term in a region larger than a single block in a grid and its semantics prefers a larger area of presence of the VisMed term to have a good match (e.g. entire kidney, skin lesion, chest x-ray images with tuberculosis), Eq. (2) will become

$$\mu\left(P,\mathrm{Z}\right)=\frac{\sum_{Z_j\in R}T_i(Z_j)}{|R|} \tag{3}$$

where Z_j are the blocks in a grid that cover *R* and |*R*| denotes the number of such blocks. This corresponds to a spatial universal quantifier (∀).

On the other hand, if a query only requires the presence of a VisMed term within a region regardless of the area of the presence (e.g. presence of a bone fracture, presence of micro nodules), then the semantics is equivalent to the spatial existential quantifier (∃) and Eq. (2) will be computed as

$$\mu\left(P,\mathrm{Z}\right)=\max_{Z_j\in R}T_i(Z_j) \tag{4}$$

A QBSI query *P* can be specified as a disjunctive normal form of VQT (with or without negation). Then the query processing of query *P* for any image *Z* is to compute the truth value μ(*P*,*Z*) using appropriate logical operators using min/max fuzzy operations. The mathematical details can be found in [9].

For the query processing in ImageCLEF 2005, a query text description is manually translated into a QBSI query with the help of a visual query interface [9] that outputs an XML format to state the VisMed terms, the spatial regions, the Boolean operators, and the spatial quantifiers. As an illustration, query 02 in the Medical Image Retrieval Task "Show me x-ray images with fractures of the femur" is translated as "∀xray-bone ∈ whole ∧ ∀xray-pelvis ∈ upper ∧ ∃xray-bone-fracture ∈ whole" where "whole" and "upper" refer to the whole image and upper part of an image respectively.

4.3. Combining Similarity- and Semantics-Based Retrieval

If a query topic is represented with both query images and text description, we can combine the similarities resulting from query processing using the FlexiTile matching scheme [8] and the fuzzy matching scheme [9]. A simple scheme would be a linear combination of λ(Q_{IMG},*Z*) and μ(P,*Z*) with λ ∈ [0,1],

$$\rho(P,Q_{IMG},x) = \omega\cdot\mu(P,Z) + (1-\omega)\cdot\lambda(Q_{IMG},Z) \tag{9}$$

where ρ is the overall similarity and the optimal λ can be determined empirically using even sampling at 0.1 intervals.
Finally, after the query image processing, we obtain, for a given query image Q_{IMG}, a ranked document list D_{IMG}.

5. TEXT AND IMAGE FUSION

We have two ranked document lists D_{IMG} and D_{TXT}. In order to obtain a unique ranked document list D_Q for each query *Q*, we propose to fusion the two lists D_{TXT} and D_{IMG} using different simple strategies. As we are working on the same document collection *D*, we make the hypothesis that the absolute Relevance Status Value (RSV) should be the same in the two lists. In practice of course they differs. We have then to rescale the RSV of the two lists using a linear transformation so that the RSV of the top document is always equal to 1.

Then, we used two simple merging techniques based on RSV_{TXT} (the RSV obtained during the text processing), and RSV_{IMG} (the RSV obtained during the image processing). Thus, for each document in both ranked list, either we keep the best ranking value ($RSV := Maximum\ (RSV_{TXT}, RSV_{IMG})$), or we compute an average value ($RSV := x\ RSV_{TXT} + (1\text{-}x)\ RSV_{IMG}$, when *x* is a constant). Keeping the best value follows the hypothesis that one media (text or image) is better to answer a query. Computing the average supposes that the two media are always participating to the ranking.

In the next section we present experimental results obtained taking into account query dimensions and media fusion.

6. EXPERIMENTAL EVALUATION

For this experiment, we have used the ImageCLEFmed-2005 Collection and the MeSH[1] ontology. For the current task, the dimensions *Anatomy*, *Pathology*, and *Modality* correspond respectively to the sub trees *Anatomy*[A], *Diseases*[C] and *Analytical, Diagnostic and Therapeutic Techniques and Equipment[E]* of the MeSH ontology.

6.1. The ImageCLEFmed-2005 corpus

As part of the Cross Language Evaluation Forum (CLEF), the ImageCLEF 2005 track [5] has a Medical Image Retrieval (MedIR) task in 2005. The test collection contains 50,026 images and annotations in XML format. The majority of the annotations are in English but a significant number is also in French and German, with a few cases that do not contain any annotation at all. The 25 queries for the MedIR task have been formulated with example images and short textual descriptions.

For the text indexing part, we used the XIOTA experimental system [3]. All documents from the same language are following a parallel processing path. Documents in a given language from all collection are merged in the same indexing matrix. We used the LTC indexing scheme of the VSM for both query and document vectors. Each query language was used to query the corresponding index matrix. We finally fused all three language results by selecting only the best matching value when the same document was retrieved from several languages in the same time. Taking the maximum value between languages emphasized the language where the matching was more efficient. For the image indexing part, we used the multi-scale detection-based approach as describe above and also in our previous work [8,9].

6.2. Experimental results

Table 1. Results obtained by query text field processing.

Hypothesis	MAP (%)	Comparison with baseline (%)
H1	0.1956	+13.39
H2	0.2075	+20.28
H3	0.1463	-17.90
H4	**0.2130**	**+23.47**

The *baseline* result obtained using the VSM without taking into account query dimensions has a Mean Average Precision (MAP) of 0.1725. To carry out the multi-dimension querying, we have made four implicit hypotheses. The results obtained are presented in Table 1 where rows correspond to the hypotheses, and values correspond to the results and their variation rates compared to the *baseline*. Here we present the four hypotheses.

[1] MeSH: Medical Subject Headings: http://www.nlm.nih.gov/mesh/meshhome.html [visited on 20/10/2005]

H1: *"Relevant documents must include at least one of the three query dimensions"*. With this hypothesis, we obtain an improvement of the result about 13.39%. In this case, we supposed that all dimensions have the same importance in the query. This assumption is not always valid. Indeed, terms describing modality in the query are not discriminative (ex: a *CT*[2] can be *"an image of a liver"* or "*an image of an emphysema"*, etc.). Also, the terms describing the pathology are, sometimes, ambiguous (ex: *a fracture* can be "*a fracture of a femur"* or "*a fracture of a cranium*", etc.). So, it seems that the anatomy is the most important dimension because it is discriminative and non ambiguous. Thus, we also make the following hypotheses:

H2: *"Relevant documents must contain the anatomy, or else the pathology, or else the modality"*. With this hypothesis, we improve the result about 20.28%.

H3: *"Relevant documents must include all the three query dimensions"*. Normally, this hypothesis should outperform the result but experiments confirm the inverse (decreasing of 17.90%). After analysing the collection, we noticed that this result is due to the fact that the CLEF documents do not usually contain terms describing the modality. For this reason, we prefer the following assumption:

H4: *"Relevant documents must contain the anatomy and the pathology dimensions"*. Thus, we obtain our best result with an improvement of 23.47%.

These results confirm the importance of using dimensions during the querying process. As we could not take into account all the query dimensions and thus, relations between them through the text processing, we think that this result can be enhanced using the complementary information present in the images. Hence, we try to compute a unique ranked document list from those obtained during the text querying process and the image querying process. For this test, we propose to use the two lists that correspond to the hypotheses that perform better our results (H2 and H4). We also use the list obtained during the image querying process with a MAP of 0.921.

We try two different strategies to merge these two lists: for each document in both lists, either we keep the best ranking (Fusion-max), or we compute an average value (Fusion-Average, where *x = 0.5*). The obtained results are presented in Table 2.

Table 2. Results obtained by text and image fusion.

Hypotheses	Text	Fusion-Average	Fusion-Max
H2	0.2075	**0.2884 (+38.98%)**	0.2355 (+13.49%)
H4	0.2130	0.2806 (+31.73%)	0.2406 (+12.29%)

Results show clearly that both visual and textual participate to the ranking. It is finally very interesting to notice that this combination outperforms both text-only and image-only results by a large amount.

The results with different values of *x* are presented in Table 3. For these tests, we used the ranked document list corresponding to the hypothesis H2 that gave our best result during media fusion. The results show that the best result is obtained when *x* is equal to 0.5. These results confirm that both text and images are important and complementary to solve multimodal queries.

Table 3. Results obtained by media fusion.

x	0	0.2	0.4	**0.5**	0.6	0.8	1
MAP	0.223	0.254	0.282	**0.2884**	0.275	0.266	0.207

7. CONSLUSION AND FUTURE WORK

In this paper we have discussed a way of incorporating external knowledge, given into ontology, to resolve multi-dimensional multimodal queries. We also used simple strategies to fusion text and images. We have performed evaluation on the ImageCLEFmed-2005 collection. In particular, we found out that external knowledge can be effectively used, and it always improves performance compared to the baseline. We also found out that results obtained using text-image fusion is far better than results obtained when processing each media separately. These experimental results show that while most queries are better handled by the text query processing as most semantic information are contained in the medical text cases, both textual and visual ontology dimensions are complementary in improving the results during media fusion.

We used this approach during our participation to the CLEF-2005 campaign and we have obtained the best result [4]. The obtained results encourage us to study the use of dimensions during the documents processing and, if necessary, to set up a new multi-dimensions indexing model. We will also study the media fusion at indexing time, and perhaps introduce an "inter-media" document indexing model.

8. REFERENCES

[1] C. Alvarez, A. I. Oumohmed, M. Mignotte, J.-Y. Nie. Toward CrossLanguage and Cross-Media Image Retrieval. *Proc. of Multilingual Information Access for Text, Speech and Images.* Proceedings of the 5th Workshop of the Cross-Language Evaluation Forum. CLEF 2004, Vol. 3491 of LNCS , Springer, UK, pp. 676–687, September 2004.

[2] M. Baziz, N. Aussenac-Gilles, M. Boughanem. Désambiguisation et Expansion de Requêtes dans un SRI, Etude de l'apport des liens sémantiques. *Revue des Sciences et Technologies de l'Information (RSTI)* série ISI, Hermes, V. 8, N. 4/2003, p. 113-136, 2003.

[3] Chevallet, J.P.: X-iota: An open xml framework for IR experimentation application on multiple weighting scheme tests in a bilingual corpus. Lecture Notes in Computer Science, Vol. 3211. AIRS'04 Conference, pp263-280, Beijing, 2004.

[4] Chevallet, J.P., Lim, J.H, Radhouani, S.: Using Ontology Dimensions and Negative Expansion to solve Precise Queries in CLEF Medical Task. In CLEF Workhop, Working Notes in Cross-Language Retrieval in Image Collections Track, Vienna, Austria, 2005.

[5] Clough, P., Muller, H.: The clef cross language image retrieval track 2005. http://ir.shef.ac.uk/imageclef2005/ [visited on November 2005]

[6] Guyot, J., Radhouani, S., Falquet, G.: Ontology-based multilingual information retrieval. In CLEF Workhop, Working Notes Multilingual Track, Vienna, Austria, 2005.

[7]. Jeon, J., Lavrenko, V., and Manmatha, R. Automatic Image Annotation and Retrieval using Cross-Media Relevance Models. In Proceedings of the 26th Annual International ACM SIGIR Conferece on Computer Vision, Vol. 4, pp.97-112, 2003.

[8] J.H. Lim & J.-P. Chevallet, "VisMed: a visual vocabulary approach for medical image indexing and retrieval," in Proc. of Asia Information Retrieval Symposium, pp. 84-96, 2005.

[9] J.H. Lim & J.S. Jin, J.S., "A structured learning framework for content-based image indexing and visual query," *Multimedia Systems Journal*, 10(4): 317-331, 2005.

[10] W.-C. Lin, Y.-C. Chang, H.-H. Chen. From Text to Image: Generating Visual Query for Image Retrieval. *Proc. of Multilingual Information Access for Text, Speech and Images*. Proceedings of the 5th Workshop of the Cross-Language Evaluation Forum. CLEF 2004, Vol. 3491 of LNCS, Springer, UK, pp. 664–675, September 2004.

[11] H. Müller, A. Geissbüuhler, P. Ruch. Report on the CLEF Experiment: Combining Image and Multi-lingual Search for Medical- Springer Lecture Notes in Computer Science, 2005.

[12] Salton, G.: Automatic Text Processing : The Transformation, Analysis and Retrieval of Information by Computer. Addison-Wesley,1989.

[13] Y. Qiu, H.-P. Frei. Concept based query expansion. *In Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, New York, pp. 160–169. 1993.

[14] E. M. Voorhees. Query expansion using lexical-semantic relations. *In Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, Springer-Verlag New York, pp. 61–69, 1994.

[2] Computed Tomography