Multimodal medical information retrieval with unsupervised rank fusion

André Mourão *, Flávio Martins, João Magalhães

Departamento de Informática, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Caparica, Portugal

ABSTRACT

Modern medical information retrieval systems are paramount to manage the insurmountable quantities of clinical data. These systems empower health care experts in the diagnosis of patients and play an important role in the clinical decision process. However, the ever-growing heterogeneous information generated in medical environments poses several challenges for retrieval systems. We propose a medical information retrieval system with support for multimodal medical case-based retrieval. The system supports medical information discovery by providing multimodal search, through a novel data fusion algorithm, and term suggestions from a medical thesaurus. Our search system compared favorably to other systems in 2013 ImageCLEFMedical.

© 2014 Elsevier Ltd. All rights reserved.

1. Introduction

A medical information storage and retrieval system is a valuable tool that healthcare professionals can use when investigating medical cases. Furthermore, case-based medical retrieval systems can empower healthcare experts by allowing them to find related publications or explore cases with similar symptoms or conditions [1] in medical information repositories. MEDLINE[1] is one of the most valuable medical information resources, containing over of 21 million articles from the life sciences and biomedical domains, covering over 5600 journals in 39 different languages. The PubMed[2] search engine was created to enable easier access to this enormous amount of information and therefore, is a major tool for retrieving biomedical literature from MEDLINE. It allows keyword search on different fields such as the main text, author, and date. However, there are some alternatives that enable different data visualization techniques or provide search tools for specific domains. Lu [2] provides an exhaustive survey of web tools for biomedical literature search.

Many existing tools focus on the full-text of the article, its categories or keywords. However, a significant percentage of medical articles contain relevant images with extra valuable information that is not referenced in the text (see Fig. 1). These images are an important clinical factor in the medical domain [3,4] and can help uncover hidden visual patterns not contained in the article's text.

The goal of our approach is to use medical images to enrich textual queries and to support discovery in medical repositories. We propose a system where healthcare specialists can write textual queries (including long descriptions of the patient's condition) and provide medical images containing additional clues that would be difficult to convey in the formulation of textual queries (see Fig. 2). Medical images can represent multiple types of diagnostic exams – X-ray images, MRI, electronic microscopy – or relevant photos that provide rich visual information about the patient, such as the position of a mass on a MRI. Our system can improve retrieval performance by combining both textual and visual information in a flexible way mixing data-fusion techniques. The key feature of our medical retrieval system interface is the support for rich multi-part queries, that can be composed by free text (with assisted query formulation) and also multiple images.

The proposed multimodal medical information system is based on standard retrieval techniques and a simple, yet powerful, design that helps users build their queries with interactive query expansion. It can retrieve articles for medical cases – medical publications and case reports – or the images contained in those articles. It is focused on usability and usefulness, for both health professionals and researchers.

We designed the system to enable exploring the bulk of medical articles available in PubMed using images and text, both independently or combined, by leveraging on their implicit correlation (images to captions and article text to article images). For instance,

* Corresponding author. Tel.: +351 212948536.
E-mail addresses: a.mourao@campus.fct.unl.pt (A. Mourão),
flaviomartins@acm.org (F. Martins), jm.magalhaes@fct.unl.pt (J. Magalhães).
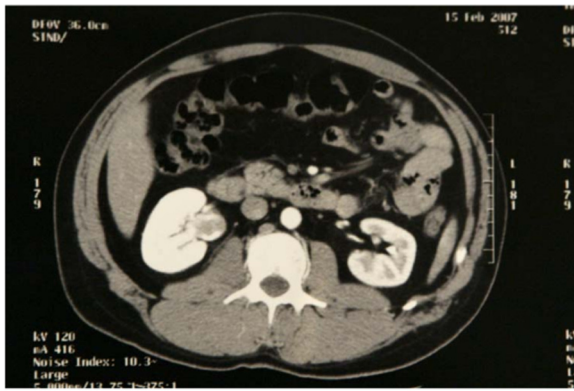
Axial computed tomography scan after intravenous adminis-
tration of iodinated contrast material, arterial phase. A filling
defect of soft tissue density is demonstrated within the right
renal pelvis, which is also dilated. Differential diagnosis
includes blood clot and tumor.

Fig. 1. Example of an image and caption from a medical article. Image adapted from: "Massive hematuria due to a congenital renal arteriovenous malformation mimicking a renal pelvis tumor: a case report" by P. Sountoulides, et al., 2008, J Med Case Rep 2 (1), page 144. Adapted with permission.
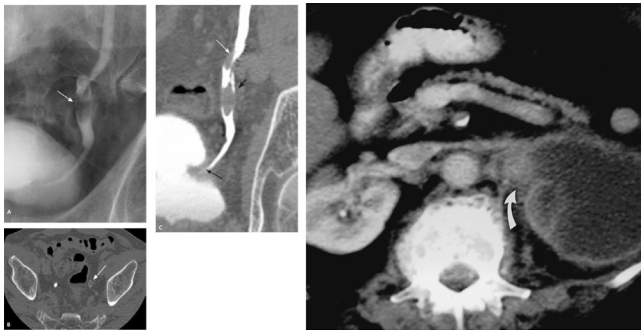
Case description:

A 43-year-old man with painless, gross hematuria.
Abdominal CT scan revealed a large left renal mass with
extension into the left renal pelvis and ureter.

Fig. 2. Multimodal medical query example from ImageCLEF 2013. It contains two images (a compound image and a CT scan) and a textual case description.

in our system, one can search for a case report using only an image as a query, or search by images using only a textual description. Besides the explicit visual information, the proposed system also provides an intuitive and simplified way of accessing large medical knowledge-bases. It identifies medical terms in real-time and suggests related terms based on the Medical Subject Headings (MeSH) thesaurus. This provides a glimpse of related conditions and diagnosis that can assist users in the formulation of a more useful query. Our system combines the simplicity of web search engines (text queries, auto-complete) with automatic query expansion and image query by simple drag and drop. Moreover, the system supports two types of visualization: medical cases (in the form of articles and case-reports) and medical images (the articles corpus is a subset of PubMedCentral, with over 70,000 articles and over 300,000 images [5]).

The system was thoroughly evaluated in the context of the 2013 ImageCLEF Medical track [3]. In this article, we start by reviewing

the state-of-the-art in medical information search in Section 2, and describing the overall architecture of the proposed system in Section 3. Details about medical terms expansion, medical images support and multimodal rank fusion are presented in Sections 4–6 respectively. Section 7 details the evaluation and discussion of the results using the most recent ImageCLEF Medical dataset and tasks: ad-hoc image retrieval and case-based retrieval tasks. Section 8 discusses our findings and conclusions.

2. Related work

2.1. Medical information search systems

Several systems designed for medical retrieval, textual and visual, are available online. Müller et al. [3] wrote a comprehensive review of content-based image retrieval systems in medical applications. The MedGIFT group [6] designed two search engine interfaces to demonstrate their work in medical information retrieval: a text based search interface[4] for cases and a visual search interface[5] for medical images. The visual search interface allows the use of uploaded images and also images from existing articles. Although these two search interfaces perform well in both modalities, they are in fact two independent systems. In other words, it is not possible to search for images using a text query or to combine images and free text to formulate a query.

img(Anaktisi)[6] [7] is a good example of a content based image retrieval (CBIR) system. It uses the Color and Edge Directivity Descriptor (CEDD) and Fuzzy Color and Texture Histogram (FCTH) image features [8] for image retrieval in multiple datasets. It is one of the few systems that includes the IRMA medical dataset and supports searching the repository using images from the dataset itself.

The Open-i search engine[7] is a recent system that supports multimodal queries to retrieve both images and articles from the PubMed Central database and on the Indiana U. Chest X-rays database. Although the system allows text and image queries, it is limited to one image per query.

In our approach, multiple images and text can be mixed and all combinations are possible. In other words, it allows searching for cases using only images, searching for images using only text or the use of both text and images in the same query for any type of result.

2.2. Medical images search and classification

Medical image retrieval systems are a widely studied retrieval problem in the medical domain, being a part of the ImageCLEF evaluation forum since 2004 [9]. Herrera et al. [10] did an overview of the latest edition, ImageCLEF Medical 2013. Images in this domain range from diagnostic images (e.g., CT scans, X-rays) to diagrams and tables included on the articles.

Most existing image retrieval systems are content-based, where the objective is to retrieve images that are visually similar to an image query. Visual similarity is represented through multiple image descriptors. The descriptors used in the medical domain are usually based on color (e.g., histograms), contours and edges (e.g., Gabor features), textures (e.g., Tamura) or other descriptors (e.g., CEDD, FCTH) that contain a combination of these. Most systems [11–15] use a combination of these descriptors. Other systems use local features based on keypoints (e.g., Scale-Invariant Feature Transform (SIFT), Speeded Up Robust Features (SURF), Oriented

---

FAST and Rotated BRIEF (ORB)) as bags of visual words. Müller et al. [3] and, more recently, Kumar et al. [16] give a more detailed review of the techniques applied in the context of medical content-based image retrieval and classification.

Medical search systems rely on standard classifiers like neural networks [11] or support vector machines [11,13] for recognizing medical image categories (e.g., X-ray, ct-scan). Kalpathy-Cramer and Hersh [11] studied the effectiveness of global image features for category classification and retrieval on ImageCLEF 2006 and 2007 datasets. They concatenated a set of features (e.g., Color Histogram, Discrete Cosine Transformation, GIST) with multiple granularities (full image, pyramid segmentation scheme) with support vector machine and neural network classifiers. They argue that global image features can be used for medical images classification with moderate success (about 80% correct detection rate) and that local features like SIFT can provide additional performance improvements. More recently, Stanley et al. [14] achieved better results (96% correct detection rate) on a different set of (nine) image categories (e.g., Chart/graph, Photograph, Radiology, Mixed) classification task with global texture features (Generalized Gray Level Spatial Dependence Models), color histograms and variations to detect if an image is mainly composed of light or dark pixels (noise). Although we can benefit from the experiments performed in these articles, it is hard to make a direct comparison with our approach, as the medical image categories and datasets are different from ImageCLEF datasets. In recent editions of ImageCLEFMedical, some teams [15,17] obtained good results with CEDD (color and image edges) and FCTH (texture and color) in the modality classification and image retrieval tasks. These descriptors are compact and contain valuable information, with fuzzy color descriptor features (e.g., blue, black) and global edge and texture image descriptors.

2.3. Medical terms expansion

Term expansion is useful to increase IR systems performance, making queries match more relevant documents that might not contain the exact query terms This effect is more visible on short queries, as suggested by Navigli and Velardi [18]. Typical query expansion techniques [19] involve taking advantage of statistical correlation, synonyms, the knowledge of the morphology of words, and the use of dictionaries to improve the quality of search results. Voorhees [20] found that automatic expansion using synonym sets from WordNet can degrade performance, while hand picking concepts improves short or ambiguous queries. Hersh et al. [21] described OHSUMED, a medical test collection with judgments, based on journals from the National Library of Medicine (NLM). Later in [22], Hersh et al. expanded the query with child terms from the Medical Subject Headings (MeSH) thesaurus, because journals in OHSUMED are indexed using the narrowest indexing terms. They also found that simply adding the new expansions to the query is a simple yet effective approach.

More recent research proposes broader techniques to cover more formal descriptions of domain-specific thesauri. This field has been quite active as is described by Bhogal et al. [23] who reviewed several ontology-based query expansion techniques. A standard model for the representation of these ontologies is the Simple Knowledge Organization System (SKOS) [24]. This standard can be used for sharing and linking controlled vocabularies (thesauri, taxonomies, classification systems, etc.) on the Web.

When expanding a term, the relationships among terms also play a key role in the process. Bai et al. [25] confirmed the importance of term-relationships expansion, and also proposed to expand queries based on the relationship between a set of terms and a single term. Thus, SKOS also provides multiple types of relationships to better assist in the query-expansion process: the *skos:altLabel* property is used to declare additional synonyms,

abbreviations and acronyms for the concept. In addition, the property *skos:related* can be used to assert non-hierarchical, associative relationships between two concepts. Fig. 6, illustrates an expanded query with multiple types of expansions.

Other approaches have mastered several techniques for integrating thesaurus-like resources in query expansion. Lin and Demner-Fushman [26] and Zhou et al. [27] investigated the creation of custom similarities to score concept matches with more emphasis than word matches.

Regarding the user interface for query expansion, there are two main approaches: AQE and IQE. Automatic query expansion (AQE) [20] adds terms to the query without user intervention, where the expanded terms are synonyms or highly related terms but the user does not receive any feedback on the expanded terms. Interactive query expansion (IQE) [28] gives the user the power to decide what terms are expanded. However, it is often an interface offered after the initial query at the cost of a more complicated interaction. Our approach is a mixture of IQE and AQE. The user can visualize what terms will be added to the query and opt-out of the expansion if incorrect or not desirable.

2.4. Multimodal fusion

Multimodal fusion algorithms combine the results from single modality (i.e., text and images) retrieval algorithms. They work at three levels: feature level fusion (early fusion), decision level fusion (late fusion) and hybrid fusion (a combination of these approaches). For a detailed description of these approaches for multimodal data see Atrey et al. [29].

In early fusion methods, features are extracted from all input data and combined into a single descriptor. This combined descriptor is sent to an analyzer (classifier, indexer). An example of an early fusion method is to concatenate the feature vectors from multiple descriptors into a single descriptor. The advantages of early fusion methods are that they make it easier to explore the correlation between the multiple data modalities. The main disadvantage is that it is necessary to represent data from all modalities into the same format (e.g., feature vector) [29]. Balancing the importance and normalization across different modalities is also a challenge for early fusion.

Late fusion methods are based on combining the decisions made at the level of single modalities into a single decision. A major category of late fusion methods for retrieval is rank fusion methods, that are (mostly) unsupervised techniques that focus on document rank and score. Rank fusion algorithms can be divided into 3 categories: score-based, rank-based and voting algorithms. Fox and Shaw [30] introduced score based fusion (CombSUM, CombMAX, CombMNZ and other variants). Score-based methods use document score retrieved by search engines as the basis for fusion. In most studies [30–32], the best performing approaches are CombSUM, CombMAX or CombMNZ, as they take the document with the higher score (in the case of CombMAX) of the sum of document scores across ranks (CombSUM, CombMAX). Other approaches offer a less representative view of the document (e.g., CombMIN orders documents by lower score across ranks). For each document i, the score after fusion can be computed as:

$$\text{combSUM}(i) = \sum_{k=1}^{N(i)} S_k(i), \tag{1}$$

$$\text{combMAX}(i) = \max(S), \forall S \subset D_i, \tag{2}$$

$$\text{combMNZ}(i) = N(i) \times \text{combSUM}(i), \tag{3}$$

where $S_k(i)$ is the score of the i document on the result list k. $N(i)$ refers to the number of times a document appears on a ranks (document frequency). $N(i)$ varies between 0 (the document i does not
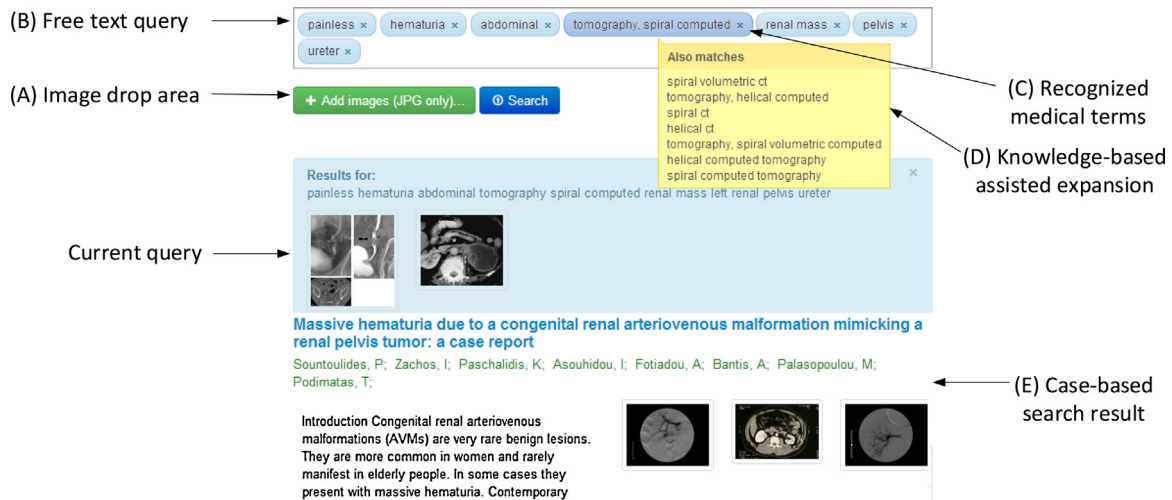
Fig. 3. NovaMedSearch interface with a multimodal query and search results.

appear on any rank) and the total number of ranks (the document i appears on all ranks). $D(i)$ are the ranks that contain the i documents.

There are some disadvantages to this approach: ranks must give documents a relevance score. Relevance scores across different modalities are computed differently and the combination of these scores become a non-trivial task, given its influence in the overall performance [33].

Rank based approaches are inspired by score based approaches, but use a rank-based virtual score (e.g., inverse of the rank). Rank based approaches can outperform score based and learning to rank methods [32,34,35] and can also be deployed in more use-cases where the score is not available. Rank based fusion methods consider the inverse of the rank of each document in each one of the individual lists as the score. Reciprocal Rank [36] and Reciprocal Rank Fusion [32] are the most well known approaches:

$$RR(i) = \sum_{k=1}^{N(i)} \frac{1}{R_k(i)}, \qquad (4)$$

$$RRF(i) = \sum_{k=1}^{N(i)} \frac{1}{h + R_k(i)}, \text{ with } h = 60, \qquad (5)$$

where $R_k(i)$ is the rank of document i on the k rank. In contrast to the score-based techniques, these techniques do not require normalization and are more stable.

The algorithms of voting approaches are based on election theory and consider the individual ranks as votes. The most well known approaches are CondorFuse [31], based on Condorcet voting and Bordafuse [37]. As rank-based approaches, these algorithms only require the ranks but are computationally more complex.
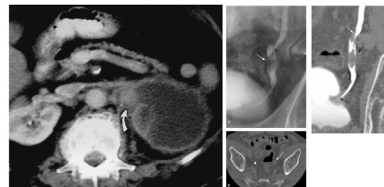
3. NovaMedSearch: multimodal medical information search

NovaMedSearch[8] (Fig. 3) is a multimodal (text and image) medical search engine that can retrieve either medical images or cases. It builds on advanced retrieval techniques and a simple design that helps users build their queries with interactive query expansion. To support rich search in the medical domain, the system provides health professionals with enough tools to best capture the

information need of the expert. Multimodal queries can contain a textual description of the case, an image textual description, and a set of relevant query images. The top section of Fig. 4 contains an example of a case description query and the corresponding images. In this figure the query text was parsed to detect medical terms and by leveraging a medical thesaurus, alternative terms are suggested on-the-fly (terms *ct-scan* and *computed tomography scan* are both recognized and expanded). The bottom section of the figure highlights how query expansion helped matching the query term ("CT scan") to a document term ("computed tomography scan").

Users can also submit multiple images in the same query. Our system supports multiple text indexes for each part of the documents (e.g., corpus, title, abstract, captions). This creates multiple

Fig. 4. Matching a query to a document example with terms expansion: The highlighted terms at the bottom represent the relevant medical terms that are matched. In yellow, we see an expanded term and how it enable matching a term in a query to the document. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)
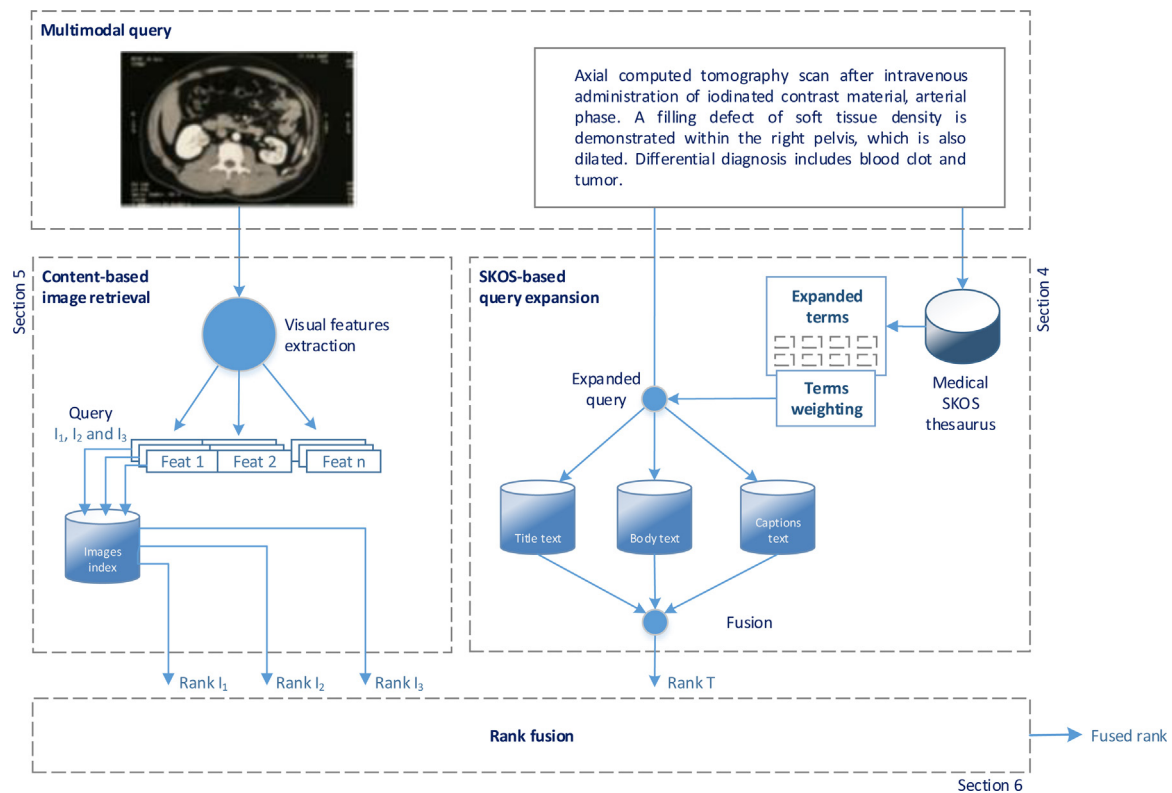

Fig. 5. Information fusion steps and the terms expansion.

candidate ranks, thus the fusion of multiple ranks plays a critical role in the overall system.

The NovaMedSearch retrieval architecture is designed to enable exploring the bulk of medical articles available at PubMed using images and text, both independently or combined, by leveraging on their implicit correlation (images to captions and article text to images). The key components of its architecture are:

- **Text retrieval**. The retrieval function BM25L [38] is at the heart of the text retrieval functionality. We experimentally found it to be the best model in the medical domain [39], in particular because it handles long documents better than other retrieval functions. A pseudo-relevance feedback technique was also implemented with the top 25 terms extracted from the top 3 documents. These features were implemented using the Apache Lucene[9] library.
- **Medical terms expansion.** We implemented a MeSH terms expansion module that is aimed at bridging terminology between the query and the indexed documents. It was integrated into Lucene with the plugin [10]. This module is described in Section 4.
- **Medical images retrieval.** Images are analysed and indexed together with their captions. The medical terms-expansion module further improves the images' semantics by finding related terms. This module is described in Section 5.
- **Multimodal ranks fusion.** The different modules produce several different ranks. The rank fusion process is based on a novel multimodal rank fusion technique, detailed in Section 6.

In the remaining of this section we introduce the main contributions of this paper: the medical terms expansion support in the user interface and the multimodal ranks fusion.

3.1. User interface: assisting medical queries formulation

The NovaMedSearch user interface, Fig. 3, aims at simplifying the inclusion of images and text data in the medical query. For instance, we add support for drag-and-drop functionality for custom medical image queries (label A in Fig. 3). The free text query box (label B) allows entering a textual description of the patient condition, and the system automatically recognizes specific medical terms (label C). For each term the user is offered a set of related term expansions – in the example, we see the terms that are related to the search term "spiral computed tomography" (label D). The search results are displayed in a ranked list with basic information (e.g., title, keywords, images) and a link to the corresponding article details (label E). The results provide a visual presentation of the submitted query and the retrieved examples. In addition to the general article information (title with link to full article, authors and abstract), we also display the images of the article that are most related to the query images.

The main novelty of the search interface is the assisted query semantic-expansion. Since medical terminology is part of natural language, terms are not unique, and multiple definitions of the same symptom/medication/disease are available. For example, our system returns "acetylsalicylic acid" and "2-(acetyloxy)benzoic acid" as terms related to "aspirin". An example for "spiral computed tomography" is in Fig. 3

We implemented a guided query expansion system that interactively provides auto-complete suggestions and expansion feedback sourced from a SKOS version of the MeSH indexing terms. Medical SKOS provides domain specific expert knowledge regarding the relationships between terms. We decided to use a SKOS version of MeSH to provide two functionalities: (1) word based auto-completion with terms, and (2) automatic term expansion with semantically related terms.
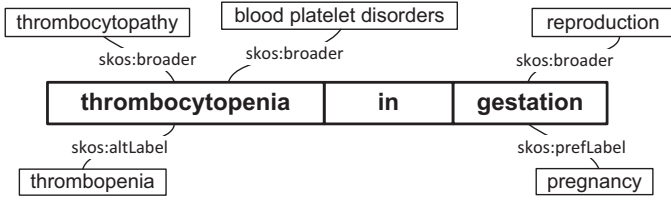
The process works as follows:

Fig. 6. SKOS query expansion example.

1. when the user starts typing a word, a dropdown box appears with the terms that match the query;
2. if the user selects a term from the list, the browser retrieves the synonyms from our framework and adds them to the query implicitly;
3. the user can then see the expanded terms by putting the mouse over the words. The user can opt-out from all suggestions by clicking the ⊗ mark.

Our system uses a SKOS representation for the terms expansion process, which makes it easy to support other Medical ontologies such as SNOMED.

3.2. Multimodal ranks fusion

Rank fusion aims at combining ranked lists from multiple sources into a single combined ranked list. Our system combines the ranks produced by the multiple query images and by searching the multiple document fields (captions, title, corpus, abstract) with the text query. A set of different retrieval techniques are used to create each rank of candidate documents. Each one of these retrieval techniques use a particular set of document characteristics to rank the documents: individual image descriptors, image modality, textual similarity to the caption, full article, etc.

Fig. 5 shows how the fusion components are connected. Image descriptors go through an early fusion process (i.e., the descriptors are concatenated). Query expansion and pseudo-relevance feedback terms are added to the query through weighted early fusion (i.e., words are searched in the same index with different weights). In this case, a late fusion is implemented within the Lucene framework producing one single text rank.

Query images can represent the same concept and belong to different categories (e.g., hepatolenticular degeneration images can be represented by a photo of an eye or by a light microscopy of the affected cells). We took a late fusion approach to combine the results from multiple image queries into a single image result list. We found that late fusion of the results was also useful for heterogeneous queries (e.g., text only, single image, text and 3 images), as the combination of the image and text search can be ignored if the query does not contain images. This final fusion produces the final rank.

4. Medical terms expansion

An information retrieval system supporting SKOS-based query expansion loads SKOS vocabularies and expands incoming queries based on the term definitions and semantic relationship in these vocabularies. Fig. 6 shows an example query "thrombocytopenia in gestation" expanded by MeSH definitions. The system found that "gestation" and "pregnancy", as well as "thrombocytopenia" and "thrombopenia" are defined as being synonyms (*skos:prefLabel, skos:altLabel*). It also found that "blood platelet disorders" is a broader term (skos:broader) for "thrombocytopenia". These terms are then added to the query, causing queries that would not match previously, such as "thrombopenia in pregnancy", to match. To consider the type of expansion for scoring and ranking, it keeps

an attribute indicating the type of SKOS property that caused the expansion.

Our proposed SKOS-based query expansion approach consist of three major building blocks: *term expansion*, *scoring* and *weighting* based on term expansion types.

4.1. Term expansion

Expanding terms requires SKOS vocabularies loaded by an information retrieval system. This can be performed either by dereferencing the URIs of concept definitions and following links to related concepts (crawling) or downloading and expanding packaged vocabularies (dump load) from a given URI. In both cases the system requires a single URI to bootstrap SKOS-based query expansion.

Term expansion at query time is part of the query analysis process and performed after the query is tokenized and before any stemming or lemmatization is applied. For each token in the user query the SKOS query expander looks up possible expansions in its internal SKOS representation, which could be represented as an inverted index having SKOS labels in the dictionary and concept Uniform resource identifiers (URIs) in the posting list. Matching expansion terms are then added to the internal query representation in which each expansion term carries an attribute indicating the expansion type, as shown in the previous example. This allows domain experts to fine-tune the influence of expanded terms on scoring of query results according to the needs in their information retrieval scenario.

4.2. Scoring the expanded terms

Term expansion can lead to a large number of terms added to the original query. While this should generally increase retrieval effectiveness, it might also cause *query drift*: the expanded query does not reflect the user's initial information need anymore. Instead of applying expensive pruning methods, our proposed scoring model leverages the explicit declaration of expansion types in query representations. This allows the configuration of to what extent certain SKOS definitions contribute to the final score.

Our approach works with regular text retrieval functions, which are minimally modified to enable scoring based on expansion types. The following function shows our scoring method as applied to the tf-idf retrieval model. The score of a query q in relation to a given document d can be succinctly conveyed as follows:

$$Score(q, d) = coord_{q,d} \cdot \left(\sum_{t \in q} (tf_{t,d} \cdot idf_t) + \sum_{c \in expansions(q)} (tf_{c,d} \cdot idf_c \cdot boost_{c_{type}}) \right)$$

$expansions(q)$ denotes the set of expansions found for query q, $coord_{q,d}$ is a score factor calculated based on the number of terms matching the document and $boost_{c_{type}}$ is a boost factor that varies with the expansion type of c.

With the $coord_{q,d}$ factor, we ensure that a document with more matching query terms will score higher. The score for documents that match more alternative terms, broader, narrower or related concepts resulting from the expansion of the original query will also be higher.

Since not all query terms are equally important, especially when they were not explicitly entered by the user, as it is the case with query expansion, original query terms and terms resulting from expansions have to be weighted differently. This can be controlled by the boost factor $boost_{c_{type}}$. It balances the contributions of the

original and expanded query terms to the final score and allows flexible weighting schemes based on expansion types. One could, for instance, weight *skos:related* expansion types less than synonym definitions (*skos:altLabel*). An optimal setup is a combination of boost values, one for each SKOS expansion type.

5. Medical images retrieval

Medical image categories in the ImageCLEF data, range from diagnostic images (e.g., radiology, microscopy) that are the basis of medical diagnosis, to charts and diagrams designed to help information visualization, explain a process or represent a mathematical equation. Müller et al. [40] created a representation based on a combination of existing terminologies that groups images into a multilevel hierarchy. These images are contained in medical articles and have varying degrees of importance for retrieval but they are seldom manually annotated. Therefore, we indexed the text associated to each image individually, and represent their visual content with different features. Each representation (i.e., caption and visual features) is indexed separately.

5.1. Visual features extraction

To capture some locality information of the general image, images are divided into 6×6 tiles and each tile is analysed individually. Two types of visual features are extracted from all tiles from all images in the corpus: Local Binary Patterns and HSV (Hue Saturation Value) color histogram.

Once all images are analysed, the extracted features are stored in a L_2 index for faster retrieval. The results of the image retrieval are sorted by their similarity, with the score being the L_2 distances between the query image and the result images.

5.2. Retrieval by visual similarity

For retrieval, the same features are extracted from the query images and passed onto the indexer to search for similar images. This process is done for each feature type. The system can answer a multimodal query with one single image and text in the order of 10^{-2} s (this excludes the image upload time). We would expect to lower this time with adequate optimisations.

For image retrieval, the results can be directly mapped into an image id (IRI) and returned to the user. For case-based retrieval, an additional step must be performed: the image id (IRI), must be converted into a document id (DOI) and the duplicate results must be merged to have an unique document list.

6. Inverse square rank fusion

There are two main approaches for late fusion: score based and rank based. Score based approaches (CombSUM, CombMAX and CombMNZ) combines the normalized scores given by individual searches (e.g., image search, textual search) as a basis to compute the new ranked list. The studied variant that is among the most popular techniques is CombMNZ [41]. Recently, it has been shown that ranked based fusion can outperform score based fusion under most conditions [32,34].

Our objective was to improve rank-based data fusion by leveraging on known techniques and prior work in score and rank-based techniques. We combined elements from both approaches to improve retrieval performance. The proposed technique, Inverse Square Rank (ISR), combines the inverse rank approach of RR and RRF (using the inverse of the rank as the score) with the document
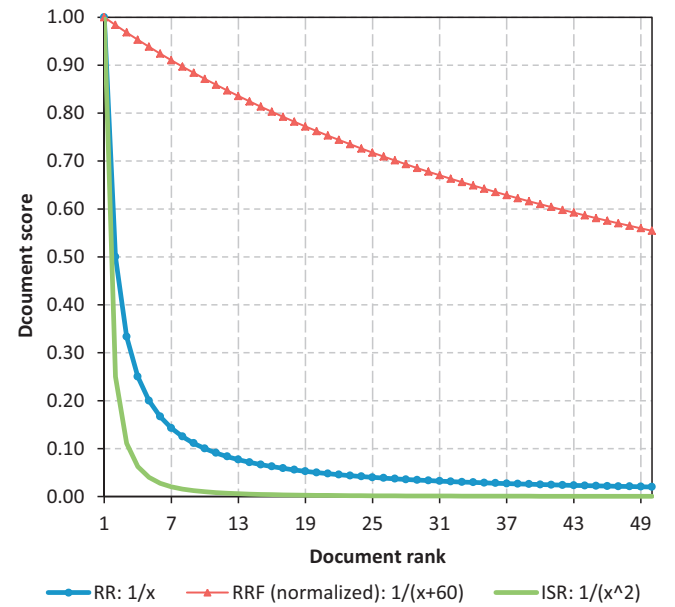

Fig. 7. Rank versus score using multiple functions.

frequency component of combMNZ (results that appear on multiple lists have higher weight). Formally, we define,

$$\mathrm{ISR}(i) = N(i) \times \sum_{k=1}^{N(i)} \frac{1}{R_k(i)^2}, \tag{6}$$

where $N(i)$ is the number of times a document appears on a results list (document frequency), and $R_k(i)$ is the rank of document i on the kth rank.

To visualize document scores assigned by ISR and other rank-based fusion algorithms, we plotted the score values for the top 50 ranked documents in Fig. 7. The RRF scores decrease slowly with the rank position. A document ranked in position 50 will have less than half of the score of the document ranked in the first position. The RR function presents a faster decay; a document ranked at position 10 has a score that is 10 times smaller than a document ranked at the first position; a document ranked 50 gets an almost negligible score. In ISR, the curve slope is even greater. This is a fundamental characteristic to merge rank lists that have high precision. Search engines are very good on the top ranked results and differences in the ranking of documents towards the end of the result list start to lose significance. By boosting the top ranked documents and multiplying by the frequency of the document across different result lists, we guarantee that documents that are highly ranked (higher probability of relevance to the query) and appear on multiple lists (to exclude documents from non-relevant engines or erratic documents) are ranked on top in the final result list.

In its simplest form, the ISR technique, weights the document-frequency using the absolute number of matched lists. We observed that linear weighting over-emphasises documents present on multiple ranks and fails to penalize documents that appear in a few ranks but are still relevant. In our initial experiments, penalizing documents that appear on a single rank, combined with logarithmic document frequency weighting, leads to a significant performance improvement. Inspired by BM25L [38] (where the logarithm was introduced to counteract increased score on long documents), we identified logarithmic normalization to provide a good model for
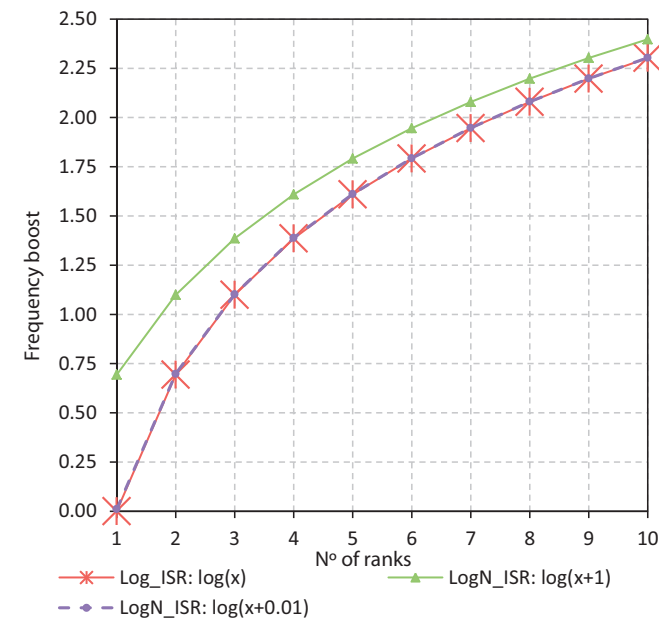
Fig. 8. Multiple document frequency weighting functions.

document frequency weighting. The final ISR techniques are log_ISR and normalized log_ISR (logN_ISR):

$$\log\_\text{ISR}(i) = \log(N(i)) \times \sum_{k=1}^{N(i)} \frac{1}{R_k(i)^2}. \tag{7}$$

$$\log N\_\text{ISR}(i) = \log(N(i) + \sigma) \times \sum_{k=1}^{N(i)} \frac{1}{R_k(i)^2}. \tag{8}$$

Fig. 8 represents the evolution of the frequency factor, using multiple normalization functions, for documents ranked from position 1–10. log_-ISR gives a score of zero to documents that appear on a single ranked list. logN_-ISR normalization factor guarantees that all documents have a *score* > 0. We tested $\sigma \in [0, 1]$, and plotted $\sigma = 0$, $\sigma = 0.01$ and $\sigma = 1$. At $\sigma = 1$ the weight difference between low frequency documents and high frequency was too small. At $\sigma = 0.01$, the weight difference is closer to the desired behavior: single rank documents are given a very small but non-zero weight. For the remaining cases, logarithmic weight grows as expected. We set $\sigma = 0.01$ for the experiments on this paper.

Log_ISR gives a score of zero to results that appear on a single rank, being dependent on a secondary ranking function. In this paper, and in line with our previous experiments, we sorted ties (documents with the same score) with a deterministic shuffle. This helps filter results from single engines that are not relevant to the query, removing "biased noise". We plan to extend our study on this phenomenon in future work.

In addition to not requiring real document scores, ISR methods can handle an arbitrary number of ranked lists with an arbitrary number of documents per list. In some tasks, some ranked lists do not contain relevant results for some queries, which decreases global retrieval performance. A further performance boost could be achieved using supervised techniques to give more weight to ranks that have the best performance or that are more adequate for a certain query. This could be applied in addition to our fusion technique, by multiplying the inverse rank score by the individual rank weight. This is out of scope for this paper, since we focused on fully unsupervised approaches.

7. Evaluation

To evaluate the proposed methods, we tested the fusion of search results on the ImageCLEFmed 2013 evaluation campaign.

7.1. Datasets

In this section, we present the dataset used and the thesaurus for query expansion.

- **Visual and Textual Medical Dataset.** This dataset is composed of the data released for medical ImageCLEF 2013. It is a subset of over 70,000 PubMed articles with over 300,000 images. Each article is identified with a unique identifier (DOI) and is divided into title, abstract, chapters and image captions. All images on the dataset have an unique identifier (IRI) and can be associated with the corresponding article and caption. We did not use any external examples or dataset augmentation in our experiments. Our training examples were the ones provided by Medical ImageCLEF 2013 [10].
- **Medical Thesauri.** The Medical Subject Headings (MeSH) is a controlled vocabulary managed by the U.S. National Library of Medicine (NLM) and is used to index millions of articles in the MEDLINE database. Physicians and medical librarians can use the terms in this vocabulary in their search activities to find the most relevant documents. MeSH is structured in various levels establishing thesaurus-like hierarchical relationships. Van Assem et al. [42] made available a SKOS version of the MeSH thesauri.

7.2. Estimation of the query-expansion parameters

To estimate the term expansion weighting described in Section 4.2 we used the TREC-9 Filtering dataset. It is a subset of the MEDLINE database with about 350,000 references from 270 journals covering four years from 1987 to 1991. Each reference contains common bibliographic metadata, including record fields for title, authors and abstract. In addition, each document contains indexing terms from the Medical Subject Headings. The collection includes three-level relevance judgments, which were assigned to documents by human assessors. The TREC-9 Filtering dataset contains

Table 1
Case-based runs performance comparison by retrieval type. All our runs and the best runs are listed. If ours is the best in the modality, the second best is also listed.

Run name	Type	MAP	GM-MAP	bpref	P@10	P@30
SNUMedinfo9	T	**0.2429**	0.1163	**0.2417**	**0.2657**	**0.1981**
FCT_LUCENE_BM25L_MSH_PRF	T	0.2233	**0.1177**	0.2000	0.2600	0.1800
FCT_LUCENE_BM25L_PRF	T	0.1992	0.0964	0.1874	0.2343	0.1781
FCT_SEGHIST_6×6_LBP	V	**0.0281**	**0.0009**	**0.0300**	**0.0429**	**0.0238**
medgift_visual_nofilter_casebased	V	0.0029	0.0001	0.0035	0.0086	0.0067
FCT_CB_MM_rComb	M	**0.1608**	0.0779	**0.1426**	0.1800	0.1257
medgift_mcasefusionandtextixed_nofilter_casebased	M	0.1467	**0.0883**	0.1318	**0.1971**	**0.1457**
FCT_CB_MM_MNZ	M	0.0794	0.0035	0.0800	0.1371	0.0810

Bold run names represent our system runs. Bold result values represent the best value for that metric.

Table 2
Medical case-based retrieval multimodal results.

Fusion	MAP	bpref	P@10	P@30
ISR	**0.1608**	**0.1426**	**0.1800**	0.1257
LogN_ISR	0.1599	0.1399	0.1657	**0.1286**
RR	0.1583	0.1367	0.1771	0.1238
RRF	0.1505	0.1218	0.1600	0.1248
combMNZ	0.0804	0.0866	0.1429	0.0819
combSUM	0.0803	0.0857	0.1457	0.0790
Log_ISR	0.0750	0.0854	0.1171	0.0790
CondorFuse	0.0433	0.0732	0.0657	0.0448
combMAX	0.0292	0.0344	0.0457	0.0248

Bold result values represent the best value for that metric.

63 queries from the original OHSUMED dataset. Please refer to [39] for details about this process.

7.3. Experiment: case-based retrieval

On the case-based retrieval experiment, the query consists of a case description (with patient demographics, limited symptoms and test results including imaging exams). The goal is to retrieve articles including images that suit the query. There is a total of 36 queries, containing patient textual description and 2–6 images each. Text and image data is indexed and searched separately, resulting in multiple ranks per query (for the image features and for text fields). The ranks are combined using the described approaches. Document ids are based on its Digital Object Identifier (DOI), meaning that each document has the same id across ranked lists.

Table 1 contains the results for the multimodal fusion experiment on the ImageCLEF 2013 case-based retrieval task. Globally, our system succeed in this task: we achieved the best results in the Visual and Mixed runs and the second best result in the Textual runs. On the Visual runs, our combination of Segmented LBP histograms and color histograms worked well when compared to other teams (about 10 times better than the second team), but the absolute results are not satisfactory (MAP and precision at 10 and 30 below 5%). On the Textual runs, results were better in general. The MeSH expansion was applied with a weigh of 1 to the original query terms and a weigh of 0.7 to the expanded terms. The runs with expansion are marked with _MSH_ in the name. Compared with the run without expansion, the run with MeSH expansion improved all metrics, especially MAP and P@10, meaning that more relevant documents are being ranked higher. On the Mixed runs, we submitted two runs: one with ISR fusion and other with combMNZ fusion, to compare the performance of score and rank based rank fusion approaches.

The medGIFT group [43] obtained the second best result in the Visual and Mixed runs. The main goal of their Visual runs was to study the effect of image category filtering in image retrieval. Their image descriptor was a combination of multiple features (CEDD + BoVW + FCTH + BoC + FCH). The query images were classified into a set of possible categories. The system retrieved a rank of candidate images, that was reranked to put images belonging to one of the candidate modalities at the top of the rank. Their results
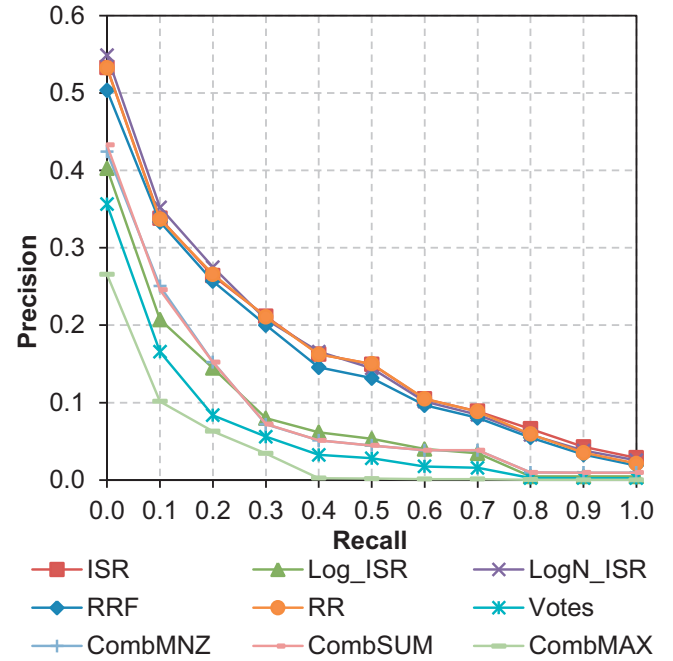
Fig. 9. Precision-recall curve for the medical case-based retrieval.

with reranking showed no significant differences to the original (non reranked) results. Compared to our Visual approach, the main difference is the descriptor set. For image and case-based retrieval, we used a concatenated descriptor with Segmented (6×6 grid) HSV histogram and Segmented (6×6 grid) Local Binary Pattern histograms.

The SNUMedinfo group [44] obtained the best result in the Textual runs by a small margin. Their text processing techniques are similar to ours (they indexed the title, abstract and full text, removed stopwords and applied stemming), but they applied pseudo-relevance feedback with data from outside the Medical ImageCLEF2013 dataset.

We performed additional experiments with other fusion functions (see Table 2). The rank-based approaches (RR, RRF, ISR, LogN_ISR) double the performance of the score-based approaches. RR obtains good performance on most metrics, but it loses on MAP to the remaining techniques. RR slightly outperforms RRF on most metrics, and on P@10 where RRF shows a significant loss. Document score decreases more slowly with rank than with RR or ISR, leading to less than optimal scoring in the top ten documents. LogN_ISR improves RRF marginally, except for a statistically relevant improvement in bpref. ISR is the best method on most metrics. It obtained the best MAP and P@10, being second only to LogN_ISR in terms of P@30.

Score-based approaches (combSUM, combMNZ, combMAX) do not perform well on this task. The reason could be the instability of the scores and the distribution differences between the cross-media ranks. CondorFuse performance was also poor. Log_ISR

Table 3
Ad-hoc image based runs performance comparison by retrieval type. All our runs and the best runs are presented.

Run Name	Type	MAP	GM-MAP	bpref	P@10	P@30
nlm-se-image-based-textual	T	**0.3196**	**0.1018**	**0.2982**	**0.3886**	**0.2686**
FCT_SOLR_BM25L_MSH	T	0.2305	0.0482	0.2316	0.2971	0.2181
FCT_SOLR_BM25L	T	0.22	0.0476	0.228	0.2657	0.2114
DEMIR4	V	**0.0185**	**0.0005**	**0.0361**	**0.0629**	**0.0581**
FCT_SEGHIST_6x6_LBP	V	0.0072	0.0001	0.0151	0.0343	0.0267

Bold run names represent our system runs. Bold result values represent the best value for that metric.

Table 4
Medical image retrieval fusion results for different fusion methods.

Fusion	MAP	bpref	P@10	P@30
RRF	**0.1508**	**0.1557**	**0.2171**	**0.1543**
LogN_ISR	0.1482	0.1529	0.2143	**0.1543**
ISR	0.1458	0.1505	0.2057	0.1476
RR	0.1450	0.1493	0.2086	0.1457
combMMZ	0.0489	0.0739	0.1457	0.0752
combSUM	0.0480	0.0730	0.1343	0.0733
Log_ISR	0.0475	0.0924	0.1400	0.0648
combMAX	0.0246	0.0317	0.0629	0.0400
CondorFuse	0.0182	0.0493	0.0314	0.0438

Bold result values represent the best value for that metric.

performs in line with score-based approaches due to the implicit instability of this function when the number of ranks to combine is small. Only the results that occur on more than one list are ranked with the expected precision. In this task, the frequency of the documents across lists is a key factor. ISR boosts results that appear on the two lists linearly, leading to the better performance.

The results are worse than in the Textual runs due to the performance difference between Text and Image runs (about 8-fold). As our approaches are unsupervised, the results retrieved by both individual modalities have the same weight on the final combined rank. A possible solution would be to learn the weights for each search engine that increase global performance on a training dataset and multiply the scores (or inverse ranks) from a search engine on the testing system. We plan to study these supervised approaches in future work.

7.4. Experiment: image retrieval

On the ad-hoc image retrieval experiment, the query consists of 1–7 sample images and a few, e.g., "thyroid CT images". The goal is to retrieve images that suit the query. There are a total of 35 queries. The rank fusion is similar to the previous experiment: text and images are indexed and searched separately, resulting in several ranks per query (one per image and another for text search

(which was fused with an internal Lucene method)). The ranks are combined using the proposed fusion methods.

Table 3 contains the results for the ad-hoc image retrieval task on ImageCLEF 2013 Medical ad-hoc image retrieval task. Our global results allow us to draw the same conclusion regarding the terms expansion method: compared with the run without expansion, the run with MeSH expansion improved all metrics, especially MAP and P@10, meaning that more relevant documents are being ranked higher.

Although we did not submit Mixed runs for the ad-hoc image retrieval task, the results of the Mixed runs with the discussed rank fusion techniques are shown in Table 4. The relative performance of the fusion algorithms is similar to the case-based retrieval. In this task, RRF slightly outperformed other approaches; we think that is because document frequency is less important here than on the case-based retrieval. LogN_ISR is the second best method. ISR and RR come in third and fourth place respectively. Score based approaches and Log_ISR again, do not perform well, being outperformed by rank-based approaches.

The precision-recall curves for case-based and image retrieval are presented in Figs. 9 and 10, respectively. The precision-recall curves follow the global results closely. ISR, logN_ISR, RRF and RR have roughly the same curve. The remaining algorithms' curves are located below. On all algorithms, precision drops sharply at 10% recall, meaning that the rank fusion methods are better at ranking the top positions.

8. Conclusions

In this article, we described our multimodal medical search engine NovaMedSearch, the technology and retrieval methods powering it, and the results and evaluation process of this system on the 2013 ImageCLEF Medical track.

The main conclusions to be drawn from our contributions concern four main points: the ISR fusion family of algorithms, an interactive query expansion system and NovaMedSearch, a search engine for the medical domain, with a special focus on usability.

The ISR algorithm family is a variant of RR and RRF, aimed at increasing precision at the top of the search engine ranks. We believe that it will help users to get relevant information, reducing frustration.

NovaMedSearch combines a powerful framework based on state-of-the-art image and text processing algorithms with a simple yet powerful multimodal search interface with interactive query expansion. Our interactive query expansion system augments user queries with relevant medical terms from the MeSH thesaurus, leveraging on the the explicit declaration of expansion types in query representations for weighting the expanded terms.

The framework that combines all search engine components was evaluated on ImageCLEF Medical 2013. Our runs achieved the best results for multimodal case based retrieval, by combining state-of-the-art image and textual retrieval techniques with our novel fusion algorithm family of algorithms. Our Image runs achieved the best results at case-based image retrieval by taking a late fusion approach on multi-image queries and using the articles where the images are present for duplicate detection. Our Textual runs also achieved good results, with a visible improvement after applying pseudo-relevance feedback and weighted query expansion with synonym terms from MeSH.
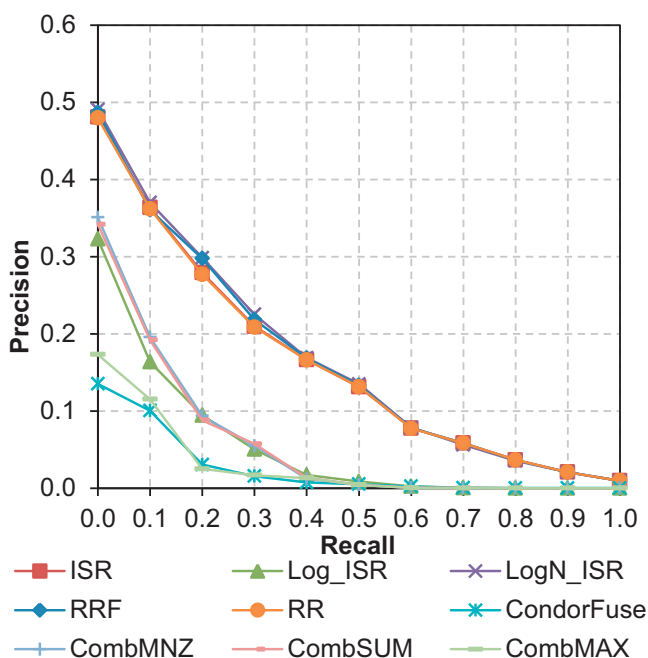
Fig. 10. Precision-recall curve for the medical image retrieval experiment.

References

[1] Samwald M, Kritz M, Gschwandtner M, Stefanov V, Hanbury A. Physicians searching the web for medical question answering: a european survey and local user studies. In: Lehmann CU, Ammenwerth E, Nøhr C, editors. Proceedings of MedInfo 2013. IOS Press; 2013. p. 1103. URL http://ebooks.iospress.nl/publication/34319

[2] Lu Z. Pubmed and beyond: a survey of web tools for searching biomedical literature. Database J Biol Databases Curation 2011;2011:1–13 http://database.oxfordjournals.org/content/2011/baq036.abstract

[3] Müller H, Michoux N, Bandon D, Geissbuhler A. A review of content-based image retrieval systems in medical applications-clinical benefits and future directions. Int J Med Inf 2004;73(1):1–23, http://dx.doi.org/10.1016/j.ijmedinf.2003.11.024 http://www.sciencedirect.com/science/article/pii/S1386505603002119

[4] Vannier M, Staab E, Clarke L. Medical image archives – present and future. In: Lemke H, Inamura K, Doi K, Vannier M, Farman A, Reiber J, editors. CARS 2002 computer assisted radiology and surgery. Berlin Heidelberg: Springer; 2002. p. 565–70, http://dx.doi.org/10.1007/978-3-642-56168-9_94.

[5] Müller H, García Seco de Herrera A, Kalpathy-Cramer J, Demner-Fushman D, Antani S, Eggel I. Overview of the imageclef 2012 medical image retrieval and classification tasks. In: CLEF 2012 Online Working Notes/Labs/Workshop. 2012.

[6] MedGIFT Group, medSearch – Medical search engine by HES-SO Valais (medSearch 2009); 2009. URL http://fast.hevs.ch:8080/MedSearch/faces/Search.jsp

[7] Zagoris K, Chatzichristos SA, Papamarkos N, Boutalis YS. A web content based image retrieval system. In: IEEE, 2009 Second International Workshop on Similarity Search and Applications. 2009. p. 154–5, http://dx.doi.org/10.1109/SISAP.2009.15.

[8] Chatzichristofis SA, Zagoris K, Boutalis YS, Papamarkos N. Accurate image retrieval based on compact composite descriptors and relevance feedback information. Int J Pattern Recogn Artif Intell (IJPRAI) 2010;24(2):207–44.

[9] Clough P, Sanderson M, Müller H. The clef cross language image retrieval track (imageclef) 2004. In: CIVR. 2004. p. 243–51.

[10] García Seco de Herrera A, Kalpathy-Cramer J, Demner-Fushman D, Antani S, Müller H. Overview of the imageclef 2013 medical tasks. In: CLEF 2013 Online Working Notes/Labs/Workshop. 2013. p. 1–15. URL http://www.clef-initiative.eu/documents/71612/4e388e82-de69-47f9-bb46-4c0ae5ebef02

[11] Kalpathy-Cramer J, Hersh W. Effectiveness of global features for automatic medical image classification and retrieval: the experiences of ohsu at imageclefmed. Pattern Recogn Lett 2008;29(15):2032–8.

[12] Pourghassem H, Ghassemian H. Content-based medical image classification using a new hierarchical merging scheme. Comput Med Imaging Graph 2008;32(8):651–61, http://dx.doi.org/10.1016/j.compmedimag.2008.07.006 http://www.sciencedirect.com/science/article/pii/S0895611108000785

[13] Rahman MM, Desai BC, Bhattacharya P. Medical image retrieval with probabilistic multi-class support vector machine classifiers and adaptive similarity fusion. Comput Med Imaging Graph 2008;32(2):95–108, http://dx.doi.org/10.1016/j.compmedimag.2007.10.001 http://www.sciencedirect.com/science/article/pii/S0895611107001383

[14] Stanley RJ, De S, Demner-Fushman D, Antani S, Thoma GR. An image feature-based approach to automatically find images for application to clinical decision support. Comput Med Imaging Graph 2011;35(5):365–72.

[15] Hosseinzadeh Vahid A, Alpkocak A, Gachpaz Hamed R, Ceylan NM, Ozturkmenoglu O. Demir at imageclefmed 2012: inter-modality and intra-modality integrated combination retrieval. In: CLEF 2012 Online Working Notes/Labs/Workshop. 2012. p. 1–13.

[16] Kumar A, Kim J, Cai W, Fulham M, Feng D. Content-based medical image retrieval: a survey of applications to multidimensional and multimodality data. J Digital Imaging 2013;26(6):1025–39, http://dx.doi.org/10.1007/s10278-013-9619-2.

[17] Vanegas JA, Caicedo JC, Camargo JE, Ramos-Pollán R, González FA. Bioingenium at imageclef 2012: text and visual indexing for medical images. In: CLEF 2012 Online Working Notes/Labs/Workshop. 2012. p. 1–11.

[18] Navigli R, Velardi P. An analysis of ontology-based query expansion strategies. In: International Workshop & Tutorial on Adaptive Text Extraction and Mining held in conjunction with the 14th European Conference on Machine Learning and the 7th European Conference on Principles and Practice of. 2003. p. 42.

[19] Carpineto C, Romano G. A survey of automatic query expansion in information retrieval. ACM Comput Surv (CSUR) 2012;44(1):1.

[20] Voorhees EM. Query expansion using lexical-semantic relations. In: SIGIR '94. 1994. p. 61–9. URL http://dl.acm.org/citation.cfm?id=188490.188508

[21] Hersh W, Buckley C, Leone TJ, Hickam D. OHSUMED: an interactive retrieval evaluation and new large test collection for research. In: SIGIR '94. 1994. p. 192–201. URL http://dl.acm.org/citation.cfm?id=188490.188557

[22] Hersh W, Price S, Donohoe L. Assessing thesaurus-based query expansion using the UMLS Metathesaurus. In: Proceedings of the AMIA Symposium. 2000. p. 344.

[23] Bhogal J, Macfarlane A, Smith P. A review of ontology based query expansion. Inf Process Manag 2007;43(4):866–86.

[24] Miles A, Bechhofer S. SKOS Simple Knowledge Organization System Reference, Recommendation, W3C; 2009 http://www.w3.org/TR/skos-reference

[25] Bai J, Song D, Bruza P, Nie J-Y, Cao G. Query expansion using term relationships in language models for information retrieval. In: ACM SIGIR'05. 2005. p. 688–95.

[26] Lin J, Demner-Fushman D. The role of knowledge in conceptual retrieval. In: SIGIR '06. 2006. p. 99–106. URL http://dl.acm.org/citation.cfm?id=1148170.1148191

[27] Zhou W, Yu C, Smalheiser N, Torvik V, Hong J. Knowledge-intensive conceptual retrieval and passage extraction of biomedical literature. In: SIGIR'07. 2007. p. 99–106.

[28] Harman D. Towards interactive query expansion. In: Proceedings of the 11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '88. New York, NY, USA: ACM; 1988. p. 321–31, http://dx.doi.org/10.1145/62437.62469.

[29] Atrey P, Hossain M, El Saddik A, Kankanhalli M. Multimodal fusion for multimedia analysis: a survey. Multimedia Syst 2010;16(6):345–79, http://dx.doi.org/10.1007/s00530-010-0182-0.

[30] Shaw JA, Fox EA. Combination of multiple searches. In: The Second Text Retrieval Conference (TREC-2). 1994. p. 243–52.

[31] Montague M, Aslam JA. Condorcet fusion for improved retrieval. In: Proceedings of the eleventh international conference on Information and knowledge management, CIKM '02. 2002. p. 538–48, http://dx.doi.org/10.1145/584792.584881.

[32] Cormack GV, Clarke CLA, Buettcher S. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In: SIGIR '09. 2009. p. 758–9, http://dx.doi.org/10.1145/1571941.1572114.

[33] Montague M, Aslam JA. Relevance score normalization for metasearch. In: Proceedings of the tenth international conference on Information and knowledge management, CIKM '01. 2001. p. 427–33, http://dx.doi.org/10.1145/502585.502657.

[34] Frank Hsu D, Taksa I. Comparing rank and score combination methods for data fusion in information retrieval. Inf Retrieval 2005;8(3):449–80, http://dx.doi.org/10.1007/s10791-005-6994-4.

[35] Macdonald C, Ounis I. Voting for candidates: adapting data fusion techniques for an expert search task. In: Proceedings of the 15th ACM international conference on Information and knowledge management, CIKM '06. 2006. p. 387–96, http://dx.doi.org/10.1145/1183614.1183671.

[36] Zhang M, Song R, Lin C, Ma S, Jiang Z, Jin Y, et al. Expansion-based technologies in finding relevant and new information: Thu trec2002 novelty track experiments. In: The Proceedings of the Eleventh Text Retrieval Conference (TREC). 2002. p. 586–90.

[37] Aslam JA, Montague M. Models for metasearch. In: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '01. 2001. p. 276–84, http://dx.doi.org/10.1145/383952.384007.

[38] Lv Y, Zhai C. When documents are very long. In: BM25 fails!, SIGIR '11. 2011. p. 1103–4. URL http://dl.acm.org/citation.cfm?id=2009916.2010070

[39] Martins F, Haslhofer B, Magalhães J. Query expansion using open web-based skos vocabularies. In: ACM SIGIR Workshop on Health Search and Discovery: Helping Users and Advancing Medicine. 2013. p. 35–8.

[40] Müller H, Kalpathy-Cramer J, Demner-Fushman D, Antani S. Creating a classification of image types in the medical literature for visual categorization. In: SPIE medical imaging, Vol. 8319. 2012., http://dx.doi.org/10.1117/12.911186, 83190P–83190P-1.

[41] Belkin NJ, Kantor P, Fox EA, Shaw JA. Combining the evidence of multiple query representations for information retrieval. Inf Process Manag 1995;31(3):431–48, http://dx.doi.org/10.1016/0306-4573(94)00057-A.

[42] Van Assem M, Malaisé V, Miles A, Schreiber G. A Method to Convert Thesauri to SKOS. Semantic Web: Res Appl 2006:95–109, http://dx.doi.org/10.1007/11762256_10 http://www.springerlink.com/index/4811513015750q25.pdf

[43] de Herrera AGS, Markonis D, Schaer R, Eggel I, Müller H. The medgift group in imageclefmed 2013. In: CLEF 2013 Online Working Notes/Labs/Workshop. 2013. p. 1–12.

[44] Choi S, Lee J, Choi J. Snumedinfo at imageclef 2013: Medical retrieval task. In: CLEF 2013 Online Working Notes/Labs/Workshop. 2013. p. 1–5.