Available online at www.sciencedirect.com
ScienceDirect
Electronic Notes in Theoretical Computer Science 225 (2009) 303–317
Electronic Notes in Theoretical Computer Science
www.elsevier.com/locate/entcs

Advanced Information Retrieval

Fuji Ren[1,2]

Department of Information Science and Intelligent Systems
The University of Tokushima
Tokushima, Japan
School of Information Engineering
Beijing University of Posts and Telecommunications
Beijing, China

David B. Bracewell[3]

Department of Information Science and Intelligent Systems
The University of Tokushima
Tokushima, Japan

Abstract

In this paper we explore some of the most important areas of advanced information retrieval. In particular, we look at cross-lingual information retrieval, multimedia information retrieval and semantic-based information retrieval. Cross-lingual information retrieval deals with asking questions in one language and retrieving documents in one or more different languages. With an increasingly globalized economy, the ability to find information in other languages is becoming a necessity. Multimedia information retrieval deals with finding media other than text, i.e. music and pictures. With the explosion of digital media that is available on the Internet and present on users' computers techniques for quickly and accurately finding desired media is important. Semantic based information retrieval goes beyond classical information retrieval and uses semantic information to understand the documents and queries in order to aid retrieval. Semantic based information retrieval goes beyond standard surface information by using the concepts represented in documents and queries to improve retrieval performance.

Keywords: Information Retrieval, Cross-lingual Information Retrieval, Multimedia Retrieval, Semantics, Semantic Based Information Retrieval

1 Introduction

Since the beginning of written language, humans have been developing ways of quickly indexing and retrieving information. From the first libraries that used alphabetization in ancient Greece to the Dewey decimal system to the Internet the amount and kind of information has grown and evolved. Information Retrieval

[1] This research has been partially supported by the Ministry of Education, Culture, Sports, Science and Technology of Japan under Grant-in-Aid for Scientific Research (B), 19300029.

[2] Email:ren@is.tokushima-u.ac.jp

[3] Email:davidb@is.tokushima-u.ac.jp

doi:10.1016/j.entcs.2008.12.082

(IR) is the act of storing, searching, and retrieving information that match a user's request [34].

Until the 1950's, information retrieval was mostly a library science. In 1945, Vannevar Bush introduced his idea of the future where machines would be used to provide easy access to the libraries of the world [5]. In the 50's the first computerized retrieval systems were designed that made use of punch cards. However, a lack of computer power limited the usefulness of these systems [20].

Starting in the 70's, computers started to have enough processing power to handle information retrieval with near instant results. With the start of the Internet, information retrieval became increasingly relevant and researched. Now, most people use some type of modern information retrieval system on a daily basis, whether it be Google or some specially created system for libraries.

This paper will explore some of the more advanced areas of information retrieval. We will focus on cross-lingual information retrieval, multimedia information retrieval and semantic-based information retrieval. Cross-lingual information retrieval deals with asking questions in one language and retrieving documents in one or more different languages. With an increasingly globalized economy, the ability to find information in other languages is becoming a necessity. Multimedia information retrieval deals with finding media other than text, i.e. music and pictures. As computers are now used for storing video and audio collections, methods for quick and accurate retrieval are needed. Finally, semantic based information retrieval goes beyond classical information retrieval and uses semantic information to understand the documents and queries in order to aid retrieval.

This paper will continue as follows. First, in section 2 we will look at at cross-lingual information retrieval. Then, in section 3 we will look at information retrieval of multimedia content. Next, in section 4 semantic-based information retrieval will be examined. Finally, in section 5 concluding remarks will be made.

2 Cross-lingual Information Retrieval

One area of information retrieval that has seen a great deal of interest and has had many exciting advances made in it, is cross-lingual information retrieval or CLIR. The goal of CLIR is to allow users to make queries in one language and retrieve documents in one or more other languages. The resulting documents can then be translated into the language used for the query to allow the user to get the gist about the information retrieved. For example, a user makes a query in English about "flower arrangement" and receives documents back in Japanese about "Ikebana," which is Japanese flower arrangement.

Recently, a number of tracks and workshops have sprung up to support research in this area. TREC (Text Retrieval Conference) had a Cross-Language IR track for a few years until 2002. CLEF (Cross Language Evaluation Forum) has been running since 2000 and deals with European languages. The NTCIR (NII Test Collection for IR Systems) Project is a yearly competition in Japan that covers many topics including CLIR dealing with languages such as Japanese, English, Chinese and

Korean. All three of these have brought together the best research in the area and have shown improvements in performance and advances in technology every year.

Most systems in CLIR use some type of translation. While there exist non-translation methods, such as cognate matching [3], latent semantic indexing [19], and relevance models [35], the predominate method is still translation. As such one of the main problems in CLIR is dealing with language translation. What should be translated, how should it be translated, and how to eliminate bad translations are some of the major areas of research in CLIR. In addition how to acquire large enough amounts of translation data is also an active topic for research. Because these, even now, are the foremost problems, this section is devoted to advanced research done to alleviate them.

This section will continue as follows. First, we will look at what to translate. Then, we look at the methods used for translation. Next, we will look at methods researchers have come up with to automatically acquire resources for translation. Finally, we look at what the future holds for CLIR.

2.1 What to Translate?

The big three choices in what to translate are the query, the document or both. Query translation involves translating the query to the target language. Document translation translates the document into the source language (i.e. the language used for the query).

2.1.1 Document Translation

Document translation is typically done using a machine translation system, such as SYSTRAN [64]. McCarley [42] points out several possible advantages to document translation. The most appealing being that by translating the document there are more chances for translating a word correctly or into a synonymous form that is used in the query.

Much research done in comparing document translation to query translation has found that document translation is typically better. Oard found that in some cases document translation gave better results than query translation on TREC-6 data [48]. Chen and Gey found that document translation gave slightly better results for the CLEF 2003 test collection [12].

However, there are some problems with document translation. The main one being that machine translation is computationally expensive and in some instances impractical [8]. However, with modern computers this is becoming less of a problem, especially for smaller document collections. Other problems include the cost of machine translation systems and the lack of availability of translation systems for a wide range of language pairs.

2.1.2 Query Translation

Query translation is typically done either using a bilingual dictionary or parallel corpus approach. The main advantage of query based translation is its speed and simplicity [42]. The main problems in query translation are dealing with ambiguities and whether word or phrases should be translated.

The simplest approach is to do word-by-word translation and use co-occurrence information for disambiguation. Fukui et al. found this type of technique to work well for patent retrieval [21]. However, this approach does not typically give the best results. Gao et al. found substantial improvements using a decaying co-occurrence approach that also utilized syntactic dependency [23]. Gao and Nie found that more specialized translation models, such as NP translation gave better results on the TREC collections [22].

In addition to standard query translation there is query expansion. Query expansion extends the query words to include similar concepts to allow for better retrieval. In monolingual IR this is typically accomplished with a thesaurus. For CLIR there are two types of expansion: pre-translation and post-translation. Pre-translation expansion adds new query words from the language the query was written in. Post-translation, takes the translated query and then extends it by some means. McNamee and Mayfield showed that even with degradation in the quality of translation (i.e. poor dictionary or error prone parallel corpus) using both pre and post translation expansion greatly improve results [43].

2.1.3 Doing both Document and Query Translation

The final option is to translate both the document and the query. While this is the most expensive it also seems to yield the best results. The reason is that document translation involves translating from the target language to the source language and query translation is the opposite, from source to target. Even when training with the same data the translation quality can be vastly different as [42] found.

Doing both translations allows for systems to take advantages of the strengths of translation in both direction. McCarley found that it gave the best results [42] for French and English. Chen and Gey also found that it gave better overall results [12].

2.2 Translation Methods

The three primary sources for translation are dictionaries, parallel corpora, and machine translation systems. Document translation, for the most part, only uses machine translation. Query translation, typically, uses either dictionary based or corpus based translation.

2.2.1 Machine Translation

The Machine translation method simply uses a machine translation system to translate either the document or query. The main drawback, as mentioned earlier, is that it is computational expensive. In situations where there is a large collection of documents or when searching for documents on the web, machine translation is impractical.

2.2.2 Parallel Corpora

Between dictionary based and corpus based translation, corpus based translation typically gives much better performance, as [43] found. However, the creation of parallel corpora is complicated and quite expensive. It can be extremely difficult to find parallel corpora for certain languages or that are large enough to be of use.

Rogati and Yang used a parallel corpus and GIZA++ [49] to determine translation probabilities that can be used for query translation [55]. Their goal was to show that degradation in performance between black box commercial machine translation systems and free material used to create a transparent system is so small that it is more preferable to use the transparent system since it allows the researcher greater control. They were able to achieve good results, an average precision greater than 0.3, on most of their tests even when using a pivot language for translation.

Nie et al. introduced a probabilistic model for CLIR that incorporates parallel corpora [47]. They tested on French and English and showed comparable results to machine translation approaches. They also introduced a simple way of gathering a parallel corpus using the web, causing them to believe their approach is more flexible than machine translation.

2.2.3 Dictionary Based

Because of the cost of machine translation and the difficulties of parallel corpora, bilingual dictionaries are widely used. A bilingual dictionary is a list of words in the source language and their translation(s) in the target language. Optionally, these dictionaries have translation probabilities assigned that allow for disambiguation and weighting.

Levow et al. looked at dictionary based CLIR in great detail [36]. They made the conclusion that CLIR is more complicated than just translation and retrieval. They also believe that studies on dictionary based CLIR can help improve corpus based CLIR.

Hedlund et al. built a dictionary based system entitled UTACLIR that works on a variety of language pairs [29]. Because they deal with many European languages, UTACLIR pays special attention to compound words that are abundant in languages like Finish and German. To deal with words that could not be translated they used N-grams for partial string matching.

2.3 Acquiring Translation Resources

The main problems with both corpus based and dictionary based translation are coverage and quality. Poor quality corpora and dictionaries can greatly decrease the performance of a system [43]. Coverage relates to out of vocabulary words, or words that are not present in the dictionary or corpus. These words will have no translation, while in some languages that are related this is no problem in other language pairs such as Chinese and English this is a big problem [75]. Because of this there has been considerable research done on automatically or semi-automatically acquiring parallel corpora or bilingual lexicons.

Some of the most prominent research done in this area is by Resnik and Smith [54]. They looked at using structural information from HTML to determine pairs of bitext web pages. Their method used search engines and a web spider to determine possible document pairs. Then, they used structural information and a content based similarity measure, for when two pages have different structure, to determine correct pairs.

Utsuro et al. used bilingual news articles to mine a Japanese-English bilingual lexicon [68]. Their system acquired comparable news articles in both English and Japanese. They then used this comparable corpus to estimate a bilingual lexicon. To improve estimation of low frequency words they then re-estimated the values using a monolingual corpus. They found that using the re-estimation was able to improve the quality of the lexicon.

2.4 The Future

While CLIR has made great advances in recent years it is still a little behind monolingual retrieval. Typically, the results are not as a good as monolingual results. In addition, acquiring lexicons and parallel corpora still remain a bit of a stumbling block especially for minority languages. In the future, we can expect to see even more research exploiting the world wide web. Finally, after CLIR has reached the level of monolingual IR, there is still the problem of how to present the information to the user. Not all users will have the ability to read the documents they retrieve. Because of this, we expect to see an increased amount of research on fast and reliable machine translation.

3 Multimedia Information Retrieval

Multimedia information retrieval (MIR) involves searching for a variety of media, such as video, music and images [41]. With the growing amount of music, video, and photos on users' computers and on the Internet the need for efficiently searching for desired media is rapidly growing. This section will take a look at the history of MIR and some of the more recent research.

The earliest research on MIR was based on computer vision research [37]. Recently, researchers are moving away from feature based retrieval to content based retrieval. There is also an increased effort to make the systems more human-centered, meaning to make the systems respond more to a user's satisfaction. Many users have started using some type of MIR, through Google Video and Image Search, Altavista Audio search, etc. While not state-of-the-art, these systems are bringing MIR to the average user.

There are numerous conferences and workshops on MIR. Some of the more prominent conferences include ACM SIGMM and the International Conference on Image and Video Retrieval. In addition there are typically special tracks in multimedia conferences, computer vision conferences, etc. dealing with MIR.

Lew et al. gave two fundamental needs for MIR systems: searching and "browsing and summarizing a media collection" [37]. The methods for achieving these needs fall mainly into two categories: feature-based and category-based. Recently, category-based methods are becoming increasingly popular, because they express the semantics of the media which allows for better retrieval.

With the two needs for MIR systems in mind, this section will continue as follows. First, we will present a look at the current research being done in music retrieval. Next, we will look at the research done on image retrieval. Then, we will look at research done on video retrieval. Finally, we will talk about the future of MIR.

3.1 Music Retrieval

In the past 5 years there has been an explosion of music made available through services such as iTunes, Napster, eMusic, etc. Even the most casual user is quickly acquiring gigabytes of music data on their computers. And there is easily petabytes of available data on the Internet. Because of this, music retrieval is a hot topic.

Downie listed a number of challenges to music information retrieval including the interaction between features such as pitch and tempo [18]. In addition he pointed out that the representation scheme determines the computational costs, such as bandwidth. Byrd and Crawford said that the same methods used in text IR, such as "conflating units of meaning", are necessary for music IR [6]. They went on to say that music IR is much harder, because there is no agreed upon definition of what a unit of meaning is and segmentation is even much harder than segmenting Chinese [6]. What features (pitch, tempo, etc), how to represent them, and what is the basic unit of music are still in debate and being researched.

Another problem is the method for querying a music database. One of the increasingly standard and popular querying methods is "query by humming." This method allows users to find songs by humming a small portion of it. One of the earlier works done by Ghias et al. focused on monophonic data [25] and used pitch in the melodic track for representation. They converted user inputed data into a symbolic form based on pitch and used this form to search a database of MIDI music [25]. Pickens et al. then extended the querying technique to deal with polyphonic music data [51]. They used a language model framework for retrieval of music performed by piano and used various methods of representation.

One noticable approach to music IR is to borrow from research in text IR. The previously mentioned research by Pickens et al. used the standard text IR approach of language modeling [51]. Uitdenbogerd and Zobel built an architecture using n-grams and approximate string matching [67]. They found that using melody information was enough for practical systems and that each of the methods, n-grams and approximate string matching, worked well for certain types of music data.

Another active area of research is music filtering. This area deals with determining which music from a collection the user may enjoy. Research has been done on automatic playlist generation [52] and music recommendation [7]. Recently, work has been done by Hijikata et al. on a content-based filtering system that has a user editable profile [30]. They employed decision trees to learn profiles of users and then allow the users to edit the trees in an online environment. They used varying features such as tempo and tonality.

3.2 Image Retrieval

In the past few years digital photography has started to overtake traditional print photography. With the growing amount of digital images, it makes sense to have an easy and effective way to search for what is desired. Instead of looking through thousands or millions of photos it is more desired to just ask "Show me all the pictures of red cars" and get the desired set of images.

Image retrieval really started in the 1970s with research done by researchers in computer vision and database management [56]. In these early days and up until the last 15 years or so, the predominant method for searching was to first annotate each image in the collection with text and then use standard text IR methods, such as [11]. Recently, as with the other areas in multimedia IR, content based retrieval is being heavily researched.

Smeulders et al. broke image retrieval applications down into three categories of user views: search by association, targets the search, and category search [61]. "Search by association" is when there is no real goal except for trying to find new interesting images. "Targets the search" is when the user has a specific image or object they are looking for. "Category search" is when users just want a picture, anyone, from a category of objects, i.e. "a car picture." With these three categories in mind, the following paragraphs will take a look at some of the research done in the area in the last few years.

Corridoni et al. looked at retrieving images based on color semantics, such as warmth, accordance, contrast etc. [14]. The system allowed users to give certain color semantics and find images that match. Kato et al. developed a system that takes a sketch done by the user and finds that image and others similar [31]. Bujis and Lew developed the imagescape application that also allows for users to sketch in images and find images similar to it [4].

Natsev et al. used multiple signatures per image to help in computing the similarity between the given image and the images in the database [46]. They found that this approach found more semantically accurate results than traditional methods. Chang et al. showed that statistical learning methods help improve the performance of visual information retrieval systems [10]. They found that they needed to introduce new algorithms to deal with sparse training data and imbalance in the type of training data. Rui et al. added relevance feedback to their MARS system to allow the user to guide the system in order to improve the search results [57]. Tieu and Viola created a framework that uses many features and a boosting algorithm to learn queries in an online manner [66]. They were able to achieve good results with only a small amount of training data, because they used selective features.

3.3 Video Retrieval

Recently, television shows, movies, documentaries, etc. have become available for download from a varying number of sites. In addition digital video and home editing is becoming the norm. Video retrieval aims to help aid the user in finding the video they seek, whether it be a full video or just a scene.

Like image retrieval some of the earliest approaches where to annotate video data and use standard IR techniques. This is still being used in modern day online video systems, such as YouTube and Google. However, with growing collections that are automatically collected from broadcast or other means annotation is impossible. As such, automatic techniques are needed. Wactlar et al. created a terabyte sized video library [70]. They used automatically acquired descriptors for indexing and segmentation.

Researchers have also tried to mimic text IR techniques in the video domain. Sivic and Zisserman made analogies between text IR and video IR [60]. Their goal was to create a fast system that works as well on video as Google does on text. They used the analogy in every facet by doing such things as building a visual vocabulary and using stop list removal. They found that while there are still some problems the analogy to text IR worked well and leaves them with future research possibilities.

Video retrieval involves such tasks as content analysis and feature extraction [1]. Aslandogan and Yu also point out that one of the most important parts of video retrieval is segmentation or partitioning [1]. Zhang et al. used multiple thresholds on the same histogram to detect gradual transitions and camera breaks [74]. Gunsel et al. looked at using syntactic and semantic features for unsupervised content-based video segmentation [28].

Sebe et al. list semantic video retrieval, learning and feedback strategies and interactive retrieval some of the new techniques used [58]. The following paragraphs will cover some of the research done using these three techniques.

Naphide and Huang used a probabilistic framework to map low level features into semantic representations [45]. The semantic representations were then used for indexing, searching and retrieval. Snoek et al. developed a semantic value chain that extracts concepts from videos [62]. They used a 32 concept lexicon and were able to achieve very good performance in the 2004 TREC Video Track.

Browne and Smeaton incorporated various relevance feedback methods and used object-based interaction and ranking [2]. Yan et al. used negative pseudo-relevance feedback for the 2002 TREC Video Track [72]. They found that this approach increased performance over standard retrieval. Yan and Hauptman introduced a boosting algorithm called Co-Retrieval for determining the most useful features [71].

Gaughan et al. built a system that incorporates speech recognition and tested in an interactive environment [24]. Girgensohn et al built a system focused on the user interface and used story segmentation with both text and visual search [26]. Their system was one the best at TRECVID.

3.4 The Future

As the amount of available multimedia data continues to grow, the need for precise MIR systems will grow also. Currently, the main push across all areas of MIR is on content based retrieval, which uses the semantics of the image, video, or audio. As the underlying algorithms improve and the semantics of audio, images, and video are better understood, the precision and usefulness of the systems will also greatly improve. In addition to improving precision, user satisfaction must be taken into account. To this end, the future of MIR will be rely greatly on strides made in affective computing.

4 Semantic Based Information Retrieval

Semantic information retrieval tries to go beyond traditional methods by defining the concepts in documents and in queries to improve retrieval. In the previous section on multimedia information retrieval, we saw that there is a current trend toward content based, or semantic, retrieval. In a similar manner semantic based information retrieval is the next evolution of text IR.

Some of the earliest work on semantic based IR was done by Raphael in 1964 [53]. He built the SIR system which broke down different queries/questions into different subroutines for processing. In a similar vein to Raphael, Li et al. looked at using semantic information for learning question classifiers [38].

Researchers have been bridging research done in semantic based IR and traditional natural language processing research fields. Li et al. used multiple information resources to help measure the semantic similarity between words [39]. Varelas et al. looked at semantic similarity methods based on WordNet and how they have applications to web based information retrieval [69].

The main methods for accomplishing semantic based IR are ontologies, semantic networks, and the semantic web. Ontologies and semantic networks can bring domain specific knowledge that allows for better performance. The semantic web, which has been a big buzz word for the past years, promises to bring semantic information in the form of standardized metadata.

This section will continue as follows. First, we will take a look at how ontologies are being used in IR. Next, we will look at research that has used semantic maps or networks. Then, we will look at the semantic web. Finally, we will talk about the future of semantic based information retrieval.

4.1 Ontologies

One common form of semantic information used in information retrieval is ontologies. Ontologies represent knowledge by linking concepts together and typically results in hierarchical classification. Khan et al. used an ontology model to generate metadata for audio and found an increase in performance over traditional keyword approaches [32]. Gomez-Perez et al. used an ontology for a legal oriented information retrieval system [27]. They found that the ontology helped guide the user in selecting better query terms. Soo et al. used an ontology as domain specfic information to increase the performance of an image retrieval system [63]. Cesarano et al. used an ontology to help categorize web pages on the fly in their semantic IR system [9].

4.2 Semantic Maps and Networks

Semantic networks, which represent concepts as nodes and relations as edges in a directed graph, are a common method used for knowledge representation. They have many uses and have been used widely in semantic based IR. Cohen and Kjeldsen developed the GRANT system that used constrained spreading activation to help in the retrieval of funding sources [13]. They found that it gave a boost to recall and precision over previous systems and had a higher level of user satisfaction. Tang et al. examined self-organizing semantic overlay networks in peer-to-peer information retrieval [65]. Lin et al. examined self-organizing semantic maps [40]. They created a semantic map based on Kohonen's self organizing map algorithm and applied it to a set of documents. The information gained from the maps allowed for easy navigation of bibliographic data.

4.3 Semantic Web

The semantic web opens a realm of new possibilities for web oriented information retrieval. Shah et al. described an approach for retrieval using the semantic web [59]. They developed a prototype system that allows for users to annotate their queries with semantic information from a couple of ontologies. Using this extra information they were able to significantly increase the precision over standard text based methods. As with other semantic information, semantic web technology can help describe domain specific information that can help improve results. Mukherjea et al. used a semantic web for biomedical patents for an information retrieval and knowledge discovery system [44]. Yu et al. looked at bringing the power of the semantic web to personal information retrieval using web services [73].

One of the main problems with the semantic web is the need for annotation. However, research such as [33], [16] and [17] is working on automatic annotation methods. Dingli et al. looked at unsupervised information extraction techniques to create seed documents which are then used to bootstrap the learning process [17]. Dill et al. built the SemTag system that was designed to automatically tag large corpora with semantic information [16].

4.4 The Future

There are a few problems facing semantic based IR. The first is the availability of semantic information sources. In English, this is not so much of a problem, but in other languages like Chinese, semantic resources are still scarce. The second problem is that, typically, algorithms dealing with semantics are much slower than the standard IR algorithms. In the future, as researchers in natural language processing progress in their own research on semantics these problems may not be so big. If the semantic web is able to reach its goal and automatic annotation methods are able to work precisely then in the future there should be no reason not to use semantic based IR, at least for the web.

5 Conclusion

This paper presented a survey of some of the areas of advanced information retrieval. We focused on cross-lingual information retrieval, multimedia information retrieval and semantic-based information retrieval. These three represent some of the most active areas of research in information retrieval. All of the presented areas have made great progress and are important for the future.

However, currently IR systems are designed to achieve high recall and precision, which is of course desired, but neglect user satisfaction. As the researchers in the multimedia information retrieval field have come to find out, future systems must make user satisfaction one of their top priorities. To this end, in the future we believe that affective computing will be a necessity for all areas of information retrieval.

Researchers in both information retrieval and affective computing see this need. Picard, one of the more important people in affective computing, gave many uses for affective computing including information retrieval [50]. Dalrymple and Zweizig performed an evaluation of information retrieval systems with respect to user satisfaction [15]. This type of research and future integration of ideas from affective computing are needed to help make IR systems human-centered.

References

[1] Aslandogan, Y. A. and C. T. Yu, *Techniques and systems for image and video retrieval*, Knowledge and Data Engineering **11** (1999), pp. 56–63.

[2] Browne, P. and A. F. Smeaton, *Video information retrieval using objects and ostensive relevance feedback*, in: *SAC '04: Proceedings of the 2004 ACM symposium on Applied computing* (2004), pp. 1084–1090.

[3] Buckley, C., M. Mitra, J. A. Walz and C. Cardie, *Using clustering and superconcepts within SMART: TREC 6*, Information Processing and Management **36** (2000), pp. 109–131.

[4] Buijs, J. M. and M. S. Lew, *Visual learning of simple semantics in imagescape*, in: *Proceedings of the Third International Conference on Visual Information and Information Systems*, 2003.

[5] Bush, V., *As we may think*, Atlantic Monthly **176** (1945), pp. 101–108.

[6] Byrd, D. and T. Crawford, *Problems of music information retrieval in the real world*, Information Processing and Management: an International Journal **38** (2002), pp. 249–272.

[7] Cano, P., M. Koppenberger and N. Wack, *An industrial-strength content-based music recommendation system*, in: *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval* (2005), pp. 673–673.

[8] Carbonell, J., Y. Yang, R. Frederking, R. Brown, Y. Geng and D. Lee, *Translingual information retrieval : A comparative evaluation*, in: *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence*, 1997, pp. 708–715.

[9] Cesarano, C., A. d'Acierno and A. Picariello, *An intelligent search agent system for semantic information retrieval on the internet*, in: *WIDM '03: Proceedings of the 5th ACM international workshop on Web information and data management* (2003), pp. 111–117.

[10] Chang, E., L. Beitao, G. Wu and K. Goh, *Statistical learning for effective visual information retrieval*, in: *IEEE International Conference on Image Processing*, 2003.

[11] Chang, S. and T. Kunii, *Pictorial database systems*, IEEE Computer **14** (1981), pp. 13–21.

[12] Chen, A. and F. C. Gey, *Combining query translation and document translation in cross-language retrieval*, in: *4th Workshop of the Cross-Language Evaluation Forum*, 2004, pp. 108–121.

[13] Cohen, P. R. and R. Kjeldsen, *Information retrieval by constrained spreading activation in semantic networks*, Inf. Process. Manage. **23** (1987), pp. 255–268.

[14] Corridoni, J., A. D. Bimbo and P. Pala, *Image retrieval by color semantics*, Multimedia Systems **7** (1999), pp. 175–183.

[15] Dalrymple, P. W. and D. L. Zweizig, *Users' experience of information retrieval systems: An exploration of the relationship between search experience and affective measures*, Library and Information Science Research **14** (1992), pp. 167–81.

[16] Dill, S., N. Eiron, D. Gibson, D. Gruhl, R. Guha, A. Jhingran, T. Kanungo, S. Rajagopalan, A. Tomkins, J. A. Tomlin and J. Y. Zien, *Semtag and seeker: bootstrapping the semantic web via automated semantic annotation*, in: *WWW '03: Proceedings of the 12th international conference on World Wide Web* (2003), pp. 178–186.

[17] Dingli, A., F. Ciravegna and Y. Wilks, *Automatic semantic annotation using unsupervised information extraction and integration*, in: *Proceedings of the K-CAP 2003 Workshop on Knowledge Markup and Semantic Annotation*, 2003.

[18] Downie, J., *Music information retrieval*, Annual Review of Information Science and Technology **37** (2003), pp. 295–340.

[19] Dumais, S. T., T. A. Letsche, M. L. Littman and T. K. Landauer, *Automatic cross-language retrieval using latent semantic indexing*, in: *AAAI Spring Symposuim on Cross-Language Text and Speech Retrieval*, 1997.

[20] Flynn, R., editor, "Computer Sciences: Macmillan Science Library," Macmillan Reference USA, 2002.

[21] Fukui, M., S. Higuchi, Y. Nakatani, M. Tanaka, A. Fuji and T. Ishikawa, *Applying a hybrid query translation method to japanese/english cross-language patent retrieval*, in: *ACM SIGIR 2000 Workshop on Patent Retrieval*, 2000.

[22] Gao, J. and J.-Y. Nie, *A study of statistical models for query translation: finding a good unit of translation*, in: *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval* (2006), pp. 194–201.

[23] Gao, J., M. Zhou, J.-Y. Nie, H. He and W. Chen, *Resolving query translation ambiguity using a decaying co-occurrence model and syntactic dependence relations*, in: *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval* (2002), pp. 183–190.

[24] Gaughan, G., A. F. Smeaton, C. Gurrin, H. Lee and K. McDonald, *Design, implementation and testing of an interactive video retrieval system*, in: *MIR '03: Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval* (2003), pp. 23–30.

[25] Ghias, A., J. Logan, D. Chamberlin and B. C. Smith, *Query by humming: musical information retrieval in an audio database*, in: *MULTIMEDIA '95: Proceedings of the third ACM international conference on Multimedia* (1995), pp. 231–236.

[26] Girgensohn, A., J. Adcock, M. D. Cooper and L. Wilcox, *A synergistic approach to efficient interactive video retrieval*, in: *INTERACT 2005*, 2005, pp. 781–794.

[27] Gomez-Perez, A., F. Ortiz-Rodriguez and B. Villazon-Terrazas, *Ontology-based legal information retrieval to improve the information access in e-government*, in: *WWW '06: Proceedings of the 15th international conference on World Wide Web* (2006), pp. 1007–1008.

[28] Gunsel, B., A. Ferman and A. Tekalp, *Temporal video segmentation using unsupervised clustering and semantic object tracking*, Journal of Electronic Imaging **7** (1998), pp. 592–604.

[29] Hedlund, T., E. Airio, H. Keskustalo, R. Lehtokangas, A. Pirkola and K. Jrvelin, *Dictionary-based cross-language information retrieval: Learning experiences from clef 2000-2002*, Information Retrieval **7** (2004), pp. 99–119.

[30] Hijikata, Y., K. Iwahama and S. Nishida, *Content-based music filtering system with editable user profile*, in: *SAC '06: Proceedings of the 2006 ACM symposium on Applied computing* (2006), pp. 1050–1057.

[31] Kato, T., T. Kurita, N. Otsu and K. Hirata, *A sketch retrieval method for full color image database-query byvisual example*, in: *Proceedings of the 11th IAPR International Conference on Computer Vision and Applications*, 1992.

[32] Khan, L., D. McLeod and E. Hovy, *Retrieval effectiveness of an ontology-based model for information selection*, The VLDB Journal The International Journal on Very Large Data Bases **13** (2004), pp. 71–85.

[33] Kiryakov, A., B. Popov, I. Terziev, D. Manov and D. Ognyanoff, *Semantic annotation, indexing, and retrieval.*, J. Web Sem. **2** (2004), pp. 49–79.

[34] Korfhage, R. R., "Information Storage and Retrieval," John Wiley and Sons, 1997.

[35] Lavrenko, V., M. Choquette and W. B. Croft, *Cross-lingual relevance models*, in: *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval* (2002), pp. 175–182.

[36] Levow, G.-A., D. W. Oard and P. Resnik, *Dictionary-based techniques for cross-language information retrieval*, Information Processing and Management: an International Journal **41** (2005), pp. 523–547.

[37] Lew, M. S., N. Sebe, C. Djeraba and R. Jain, *Content-based multimedia information retrieval: State of the art and challenges*, ACM Trans. Multimedia Comput. Commun. Appl. **2** (2006), pp. 1–19.

[38] Li, X., D. Roth and K. Small, *The role of semantic information in learning question classifiers*, in: *The First International Joint Conference on Natural Language Processing*, 2004.

[39] Li, Y., Z. A. Bandar and D. McLean, *An approach for measuring semantic similarity between words using multiple information sources*, IEEE Transactions on Knowledge and Data Engineering **15** (2003), pp. 871–882.

[40] Lin, X., D. Soergel and G. Marchionini, *A self-organizing semantic map for information retrieval*, in: *SIGIR '91: Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval* (1991), pp. 262–269.

[41] Maybury, M. T., editor, "Intelligent Multimedia Information Retrieval," AAAI Press, 1997.

[42] McCarley, J. S., *Should we translate the documents or the queries in cross-language information retrieval?*, in: *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics* (1999), pp. 208–214.

[43] McNamee, P. and J. Mayfield, *Comparing cross-language query expansion techniques by degrading translation resources*, in: *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval* (2002), pp. 159–166.

[44] Mukherjea, S., B. Bamba and P. Kankar, *Information retrieval and knowledge discovery utilizing a biomedical patent semantic web*, IEEE Transactions on Knowledge and Data Engineering **17** (2005), pp. 1099–1110.

[45] Naphide, H. and T. Huang, *A probabilistic framework for semantic video indexing, filtering, and retrieval*, IEEE Transactions on Multimedia **3** (2001), pp. 141–151.

[46] Natsev, A., R. Rastogi and K. Shim, *Walrus: a similarity retrieval algorithm for image databases*, IEEE Transactions on Knowledge and Data Engineering **16** (2004), pp. 301–316.

[47] Nie, J.-Y., M. Simard, P. Isabelle and R. Durand, *Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the web*, in: *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* (1999), pp. 74–81.

[48] Oard, D. W., *A comparative study of query and document translation for cross-language information retrieval*, in: *AMTA '98: Proceedings of the Third Conference of the Association for Machine Translation in the Americas on Machine Translation and the Information Soup* (1998), pp. 472–483.

[49] Och, F. J. and H. Ney, *A systematic comparison of various statistical alignment models*, Computational Linguistics **29** (2003), pp. 19–51.

[50] Picard, R. W., "Affective Computing," MIT Press, 2000.

[51] Pickens, J., J. P. Bello, G. Mont, M. Sandler, T. Crawford, M. Dovey and D. Byrd, *Polyphonic score retrieval using polyphonic audio queries: A harmonic modeling approach*, Journal of New Music Research **32** (2003), pp. 223 – 236.

[52] Ragno, R., C. J. C. Burges and C. Herley, *Inferring similarity between music objects with application to playlist generation*, in: *MIR '05: Proceedings of the 7th ACM SIGMM international workshop on Multimedia information retrieval* (2005), pp. 73–80.

[53] Raphael, B., *Sir: A computer program for semantic information retrieval*, Technical report, Cambridge, MA, USA (1964).

[54] Resnik, P. and N. A. Smith, *The web as a parallel corpus*, Computational Linguistics **29** (2003), pp. 349–380.

[55] Rogati, M. and Y. Yang, *Multilingual information retrieval using open, transparent resources in clef 2003*, in: *CLEF 2003*, 2003, pp. 133–139.

[56] Rui, Y., T. Huang and S. Chang, *Image retrieval: current techniques, promising directions and open issues*, Journal of Visual Communication and Image Representation **10** (1999), pp. 39–62.

[57] Rui, Y., T. S. Huang, S. Mehrotra and M. Ortega, *A relevance feedback architecture for content-based multimedia information retrieval systems*, cbaivl **00** (1997), p. 82.

[58] Sebe, N., M. S. Lew, X. Zhou, T. S. Huang and E. M. Bakker, *The state of the art in image and video retrieval*, in: *Image and Video Retrieval*, 2003.

[59] Shah, U., T. Finin and A. Joshi, *Information retrieval on the semantic web*, in: *CIKM '02: Proceedings of the eleventh international conference on Information and knowledge management* (2002), pp. 461–468.

[60] Sivic, J. and A. Zisserman, *Video google: a text retrieval approach to object matching in videos*, in: *Proceedings of the Ninth IEEE International Conference on Computer Vision*, 2003.

[61] Smeulders, A. W. M., M. Worring, S. Santini, A. Gupta and R. Jain, *Content-based image retrieval at the end of the early years*, IEEE Trans. Pattern Anal. Mach. Intell. **22** (2000), pp. 1349–1380.

[62] Snoek, C. G., M. Worring, J.-M. Geusebroek, D. C. Koelma and F. J. Seinstra, *The mediamill TRECVID 2004 semantic video search engine*, in: *Proceedings of the 2th TRECVID Workshop*, 2004.

[63] Soo, V.-W., C.-Y. Lee, C.-C. Li, S. L. Chen and C. chih Chen, *Automated semantic annotation and retrieval based on sharable ontology and case-based learning techniques*, in: *JCDL '03: Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries* (2003), pp. 61–72.

[64] Systran, *Systran language translation technology*, [Online], http://www.systransoft.com.

[65] Tang, C., Z. Xu and S. Dwarkadas, *Peer-to-peer information retrieval using self-organizing semantic overlay networks*, in: *SIGCOMM '03: Proceedings of the 2003 conference on Applications, technologies, architectures, and protocols for computer communications* (2003), pp. 175–186.

[66] Tieu, K. and P. Viola, *Boosting image retrieval*, International Journal of Computer Vision **56** (2004), pp. 17–36.

[67] Uitdenbogerd, A. L. and J. Zobel, *An architecture for effective music information retrieval*, Journal of the American Society for Information Science and Technology **55** (2004), pp. 1053–1057.

[68] Utsuro, T., K. Hino, M. Kida, S. Nakagawa and S. Sato, *Integrating cross-lingually relevant news articles and monolingual web documents in bilingual lexicon acquisition*, in: *Proceedings of Coling 2004* (2004), pp. 1036–1042.

[69] Varelas, G., E. Voutsakis, P. Raftopoulou, E. G. Petrakis and E. E. Milios, *Semantic similarity methods in wordnet and their application to information retrieval on the web*, in: *WIDM '05: Proceedings of the 7th annual ACM international workshop on Web information and data management* (2005), pp. 10–16.

[70] Wactlar, H. D., M. G. Christel, Y. Gong and A. G. Hauptmann, *Lessons learned from building a terabyte digital video library*, Computer **32** (1999), pp. 66–73.

[71] Yan, R. and A. Hauptman, *Co-retrieval: a boosted reranking approach for video retrieval*, Vision, Image and Signal Processing **152** (2005), pp. 888– 895.

[72] Yan, R., A. G. Hauptmann and R. Jin, *Negative pseudo-relevance feedback in content-based video retrieval*, in: *MULTIMEDIA '03: Proceedings of the eleventh ACM international conference on Multimedia* (2003), pp. 343–346.

[73] Yu, H., T. Mine and M. Amamiya, *An architecture for personal semantic web information retrieval system–integrating web services and web contents*, in: *IEEE International Conference on Web Services*, 2005, pp. 329–336.

[74] Zhang, H., A. Kankanhalli and S. W. Smoliar, *Automatic partitioning of full-motion video*, Multimedia Syst. **1** (1993), pp. 10–28.

[75] Zhang, Y. and P. Vines, *Using the web for automated translation extraction in cross-language information retrieval*, in: *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval* (2004), pp. 162–169.