# Going Beyond the Surrounding Text to Semantically Annotate and Search Digital Images

Shahrul Azman Noah[1], Datul Aida Ali[1], Arifah Che Alhadi[2], and Junaidah Mohamad Kassim[1]

[1] Faculty of Information Science & Technology, Universiti Kebangsaan Malaysia 43600 UKM Bangi, Selangor, Malaysia
{samn,da41030,junaidah}@ftsm.ukm.my
[2] Department of Computer Science, Universiti Malaysia Terengganu 21030 Kuala Terengganu, Terengganu, Malaysia
arifah_hadi@umt.edu.my

**Abstract.** Digital objects such as images and videos are fundamental resources in digital library. To assist in retrieving such objects usually they are being tagged by some keywords or sentences. The popular approach to tag digital objects is based on associated text. However, relying on associated text alone such as the surrounding text unable to semantically describe such objects. This paper discusses the use of WordNet and ConceptNet to tag digital images beyond terms available in the surrounding text. WordNet is used to disambiguate concepts or terms from the associated text and ConceptNet is meant to infer topics or common-sense knowledge from summarizing the text that describe the images. However, relying on WordNet alone is not sufficed particularly when it comes to disambiguate specific or domain dependent concepts. As such the Name Entity Recognition (NER) technique is required to annotate important entities such as name of a person, location and organization. Our work focused on on-lines news images that are richly described with textual description.

**Keywords:** semantic annotation, information retrieval, ontology, natural language processing.

## 1 Introduction

Card catalogs existed long before computers did, which means that these cards were constructed manually. Given a book, a person had to extract the author, title and subject headings of the book so that the various catalogs could be built. This process is generally known as manual annotation. Manual annotation can also be found in the current Web 2.0 applications whereby users can collaboratively annotate and describe digital objects such as images and videos. A few efforts have been put forward to automatically annotate digital objects particularly images. In this sense automatic image annotation can be regarded as the process by which an application automatically assigns metadata in the form of captioning or keywords to a digital image. The apparent advantages of automatic image annotation versus content-based image retrieval (CBIR) are that queries can be more naturally specified by the user [1]. CBIR

N.T. Nguyen, M.T. Le, and J. Świątek (Eds.): ACIIDS 2010, Part I, LNAI 5990, pp. 169–179, 2010.

systems on the other hand typically require users to express their information needs in a very unnatural way in terms of low-level features such as colors, textures or example images.

Vast amount of images are now available as digital format and many of them are being described either directly or indirectly by some textual information [2]. Such characteristics of images are also common in many digital libraries and other information resources centers. While many approaches have mainly focused on keyword spotting for annotation and searching, little have gone beyond to the "semantic meaning" of images which require further conceptualization and summarization of its corresponding textual descriptions. Search engines for instance consider all the surrounding text or the nearby text (excluding the specified stopwords) to annotate images but leave the semantics meanings of images to the users. The use Latent Semantic Indexing (LSI) might be helpful but highly depends on the textual document collections.

This paper reports our approach in generating semantic annotations for images in order to support semantic search. Our proposed approach is most suitable for images with textual description and testing has been conducted on on-line newspaper images. Using the natural language processing (NLP) techniques, named entity recognition (NER), and lexical ontology, the main inherent idea is to extract semantic meanings of images based from the free (unstructured) textual annotation provided by web authors and then subsequently construct the semantic annotation.

## 2 Background and Related Research

In general terms, annotation means to add explanation and notes to a lot of things such as an artifacts, book and even images with the intention of giving additional information. There are two common types of annotation: structured annotation and unstructured annotation. We defined structured annotation as comments made directly to sources by means of some tools and which are usually seen technically as metadata. However, in real life there are lots of unstructured annotation that can be seen as natural language description associated to images, documents and artifacts. We called this unstructured because it is not represented by means of metadata but usually freely provided by document or web authors. Our scope of work does not cover work related to low-level image feature extraction. As mentioned earlier, automatic annotation approaches usually gather such textual description, preprocessed them and used them as annotation to images. However, such an approach is very limited in many senses. For example, not all descriptions accurately described the meaning of images and on some cases the descriptions are provided at other places in documents containing the images. More obviously in practice, annotations are usually derived from the overall content of documents and usually provided by a single or at most three words as in the cases of a human (or a librarian) assigning keywords to particular objects (or documents). Our proposed approach is two-fold: first is to semantically enhance annotations by means of referring to lexical ontologies; and second is to annotate important entities among image searchers such as 'person', 'location' and 'time' using the NER patterns so that when a query similar to '*Johan in Kuala Lumpur*' can be 'semantically' conclude that '*Johan*' is a person and '*Kuala Lumpur*' is a location.

### 2.1 WordNet and ConceptNet

WordNet is a lexical database for the English language. Its development was inspired by psycholinguistic theories of human lexical memory. WordNet groups English words into sets of synonyms called synsets, and records the various semantic relations between these synonym sets. WordNet is usually used as a dictionary and thesaurus, or as a tool to support automatic text analysis. In the area of information retrieval, WordNet is used on a number of aspect such as semantic retrieval [3], word sense disambiguation [4] and query expansion [5]. In our approach, WordNet is used to provide additional meanings of the indexed terms (based on the surrounding text) of images. In this case, based upon our collection of on-line Malaysian news images, the term 'prime minister' will be provided with additional meanings such as 'head of state', and 'chief minister'.

ConceptNet [6] on the other hand is a sophisticated multilingual lexical resource organized around meanings of words and expressions that supports many practical textual-reasoning tasks over real-world documents including topic-gisting, affect-sensing, analogy-making, and other context-oriented inferences. ConceptNet contains over 20 types of relationships between concepts (synonym, part-of, instance-of, belong-to-domain, etc.) and over 500,000 actual relationships among concept. ConceptNet is, therefore, not a dictionary or thesaurus as in the case of WordNet. It is because ConceptNet is organized in concepts that use relations to build hierarchies and networks.

In information retrieval, ConceptNet can integrate optimally with search solution and allows for true interactive cross-language [7]. This means that users can select meaning of words and expressions they want to look for. Eagle et al. [8] for instance used ConceptNet (topic-gisting) to identify conversation topic from devices like PDA or hand phones. They use semantic relationship like LocationOf, SubeventOf, HasEffect to recognize conversation topic. In our approach, ConceptNet is used to identify the suitable topic for images by using its topic-gisting feature. For example the text "*Two Law students from Universiti Malaya (UM) have won the prestigious Anugerah Pelajaran Diraja. Mohd Fatihin Awang Ali, who studies Syariah law, and Neoh Hor Kee received the award consisting of cash, a gold medal and a certificate……*" would generate the topic: '*award*', '*trophy*', '*dean*', '*competition*', '*university*' and '*surprise*'.

### 2.2 Related Work

Research in semantic annotation of digital images from surrounding text is illustrated from the work of [2, 9,10]. Benitez and Chang [2] focused on extracting the meaning and relationships from existing collection of annotated images. The derivation of meanings includes three-stage of processing: text processing; extraction of semantic concept; and extraction of semantic relationships. Text processing involves tokenizing and chunking the textual annotations (which are in the form of sentence) and assign the Part-of-Speech (POS) tag to each word. Each word is disambiguated using the WordNet in order to define its semantic concept. Weights are assigned to chosen concepts using the $tf \times idf$ and $\log tf \times entropy$ schemes. Weights are also given based on synsets, core meaning and example usage. The extraction of semantic relationships

on the other hand is to add additional concepts from the WordNet in order to relate the detected word senses. Their experiments show that WordNet gives better result for images under the category of nature. In addition to this, their experiment also shows that annotated news images that contain more textual description than those images that contain only keywords also give better result.

The work of Gong et al [9] encircle around developing and indexing scheme of web images using texts that are available in the web. This research emphasizes the principle that every website developer will use images to describe their website and the distance between the word and the image are also important. Therefore, texts that are available in the website will have semantic relations that are linked to the image. Their approach divides the text containing the images into three blocks of semantic segmentation, i.e. i) TM (Title & Meta); ii) LT (image location, image name, image hyperlink or ALT); and iii) BT (body of text). Results from the experiment show that texts situated nearest to the image gives better recall measure. However, when taking the whole text into account (i.e. by considering the three segments of TM, LT and BT), the recall measures outperform the others. The work of [9], however, does not consider external resources to provide additional semantic meanings of the indexed terms.

The work of Zhigang et al [10] combined visual and textual feature for searching images. Semantic information extracted from web pages are text summary, human related information such as name of a person, geographical information such as name of a place and telephone number. Apart from concept extraction as in the previous two approaches, their approach proposed four aspects of semantic information extraction namely: visual weight, total phrases, phrase weight and independent phrases. Result shows that 62% to 90% of web images capable of being semantically described. The proposed approach, however, still could not differentiate between geographical and human name such as McDonald's which should be considered as geographical information and not name of a person.

The previous three approaches are representative of other approaches [11,12,13,14,15,16] that use associated text for image annotations. Most approaches exploit the location of text for selecting suitable annotation terms. Few will further disambiguate these terms by referring to the WordNet. However, not all images' descriptive terms can be semantically extended by WordNet. Therefore, some common-sense concepts are required. For instance mentioning the word 'bride and groom' or seeing images with similar 'content' will relate to terms such as 'wedding' and 'ring' [17]. Such common-sense concepts can be derived from semantically processed resources such as the ConceptNet. On some cases WordNet will not be able to disambiguate very specific concepts such as name of a person, name of an organization and some unfamiliar geographical locations. In this case NER is required. This study, therefore, embarks on the possibility of assigning annotation beyond the normal keywords extracted from surrounding text by using semantic and common-sense information from the WordNet and ConceptNet. As some concepts cannot be disambiguated by WordNet, NER patterns have been developed to identify important entities such as person, location and time.

## 3 The Approach

Annotation is the process of giving meanings to some artifacts. For digital images we provide meanings that best describe the content of images. However as the old Chinese proverb says “a picture worth a thousand words”, automatic textual annotation of images proved to be difficult and relying on the surrounding text alone is not suffice. Our approach considers the use of WordNet and ConceptNet, as well as NER patterns. In this study, we scope our work into on-line Malaysian English language newspaper under the nation category. The nation category is related to Malaysian political and local issue news. Our approach is as depicted in Fig 1, consisting a number of stages as follows.

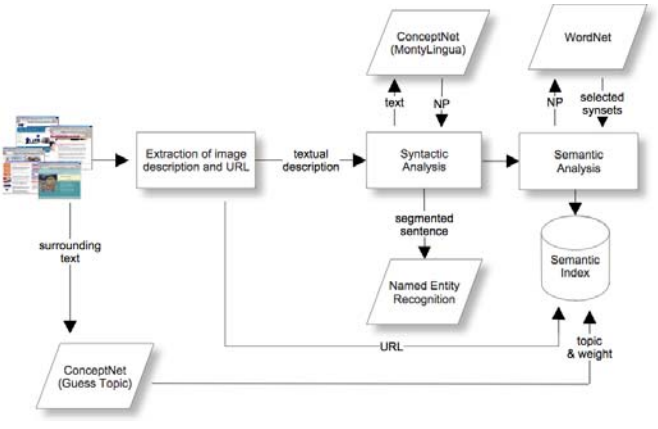


**Fig. 1.** The process to support semantic search of on-line news images

### 3.1 Extraction of Image Description, Url and Surrounding Text

As mentioned earlier, textual information surrounding images can be regarded as unstructured annotation of images. Images in on-line newspapers contain such rich and useful annotations. In this stage, tags in HTML documents have to be ‘cleaned’ with the exception of ‘p’, 'br', 'div', and 'span' tags. These tags are required as a bookmark for extracting the surrounding text, image description and image location. The HTML documents will then transformed to ASCII document file type. Image location will be stored directly into the database whereas the associated image’s description and surrounding text will be used in the next process. Throughout this paper, image’s description refers to textual information right under images as exhibited in many on-line newspaper images. Other textual information is considered as surrounding text.

### 3.2 Syntactic Analysis

Syntactic analysis contains two main stages: syntax analysis and NER. The syntax analysis involved the MontyLingua that is part of the ConceptNet component. MontyLingua is a NLP toolkit containing suite of libraries and programs for symbolic and statistical NLP. Every sentence in the description will be tokenized using

MontyTokenize classes called tokenize and tag_tokenize. The class tokenize transforms input sentences into sentences without hyphenation, whereas the class tag_tokenize performs the part-of-speech (POS) tagging process following the Penn TreeBank tagsets. The output of this stage is in the form of " *This/DT is/VB a/DT sentence/NN* ".

After that the sentence will be tagged with a more detail tag in the MontyTagger class and the output sentence will be like this: " *(NX He/PRP NX)(VX is/VB VX) (NX the/DT mailman/NN NX)* ". Only noun phrases will be considered for the next process based on the assumptions that noun phrases are the best lexical category to describe images [18]. Weight for every noun phrases are calculated using the *tf×idf* weighting scheme.

The second stage which is the NER stage contains two sub-processes: the sentence segmentation process and pattern matching process. In sentence segmentation, noun phrases from syntax analysis are used for sentence selection. Fig 2 illustrates an example of a word matching for choosing segmented sentence. In this example, the sentence: "*Friendly visit: Anwar talking to Mohammad Nizar in Ipoh yesterday. With them are (from left) Gopeng MP Dr Lee Boon Chye, Anwar's wife Datin Seri Dr Wan Azizah Wan Ismail and Perak PKR chief Zulkifly Ibrahim*", will be segmented into "Anwar talking to Mohmmad Nizar", "Mohammad Nizar in Ipoh: and "Gopeng MP Dr Lee Boon Chye" based upon the word matching of the set {"Abdullah', 'Ipoh', 'Anwar' and 'Sungai Dua'}.

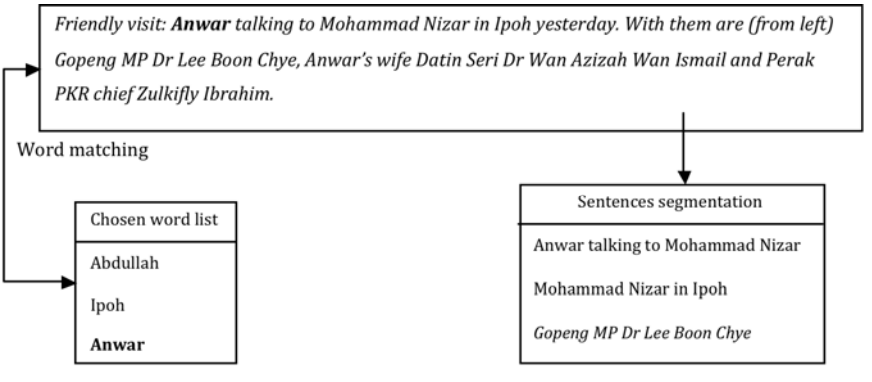


**Fig. 2.** The process of sentence segmentation

The next process of this second stage is the pattern matching process whereby the segmented sentences are matched with NER patterns developed from the analysis of existing documents. Overall there are 142 patterns meant to extract entities such as 'person', 'location', 'event', 'time', 'title', 'organization' and 'position'. These entities are chosen based on the opinion of Kawata et al. [19] that such entities are the most common information acquired in news. Table 1 lists the example of the patterns used in this study and Fig 3 and 4 illustrate the approach and an example of the pattern matching process respectively.

**Table 1.** Example of the NER pattern

| Named entity | Pattern |
|---|---|
| Person | [person] having, [person] watching her [post] [person] |
| Location | in [location], [location], [person] at the [location] in [location] |
| Event | at the [event], the [event] held at |
| Time | at the [location] on [time], in [location] since the [time] |
| Title | [post][title][person], [title][person] |
| Person-post | [post] from the [post], the [person's post] holding up |
| Organization | at the [organization], [person] from [organization] |

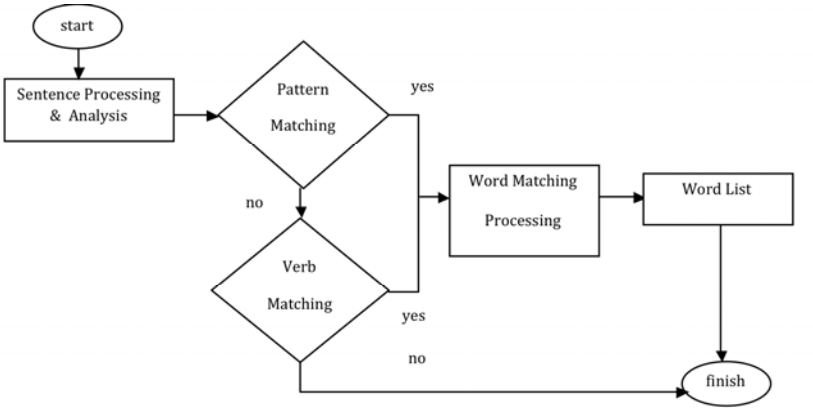


**Fig. 3.** The flow of a pattern matching process

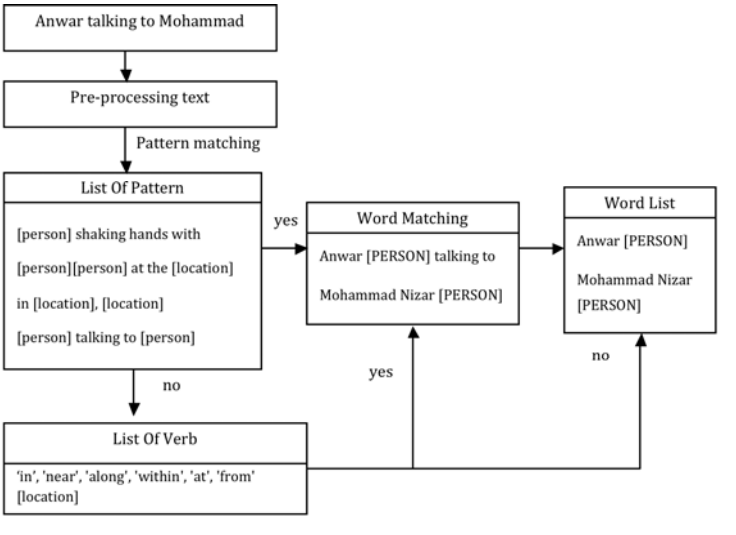


**Fig. 4.** An example of a pattern matching process

Based on Fig 4, the input for pattern matching are the segmented sentences. The sentences will be parsed to generate the POS tags (preprocessing text). For instance, the sentence "*Anwar talking to Mohammad Nizar*" will be tagged and tokenized as *"(NX Anwar/NX)", "(VX talking to/VX)", "(NX Mohammad Nizar/NX)*". The tagged sentence is then matched with the appropriate pattern. For this example, it will be matched with the pattern *"[person] talking to [person]*" and subsequently semantically tagged as *"(Anwar [PERSON]) talking to (Mohammad Nizar [PERSON])*". For those noun phrases in the word list that do not matched with any of the patterns, the process will continue by matching the verbs with the NER patterns. For example the tokenize segmented sentence of (*NX SK St Teresa/NX) (VXalong/VX)(NX Jalan Brickfields/NX))* will resulted in the generation of "(*SK St Teresa[LOCATION]) along (Jalan Brickfields [LOCATION])*".

### 3.3 Deriving Semantic Meanings

The main aim of this stage is to add additional semantic information and common sense knowledge of the annotated images using the WordNet and ConceptNet. Noun phrases from the word list are matched with the terms in WordNet. If such matched exists, the hypernyms of the terms will be included as part of the annotated terms. For instance, the phrase "*prime minister*" will generate semantic information such as "*Head of State, Chief of State, Representative, and Negotiator*.

As mentioned earlier, ConceptNet is used to derive common-sense knowledge or meanings of sentences. Such meanings can be derived using the topic-gisting module. To achieve this, sentences are first fragmented into verb-subject-object-object (VSOO), of which then used to derive related concepts (or topics) from the ConceptNet. ConceptNet provides weights (saliency weight) based on the relevancy of the topics to the submitted sentences. These weights are based upon lightweight syntactic cues and contextual intersection. It is not realistic, however, to consider all the topics derived from the ConceptNet. Therefore, we consider only top topics, i.e. those top $n$ topics before the weight has converged to a specific $\delta$ value. The ConceptNet topic gisting module allows terms (or topics) associated with the images (based on the description or surrounding text) to be derived from a large lexical resource. Liu & Lieberman [20] refer these additional terms as common sense knowledge.

## 4 Evaluation

Initial evaluation was conducted by comparing our proposed approach to semantic search with the conventional bag-of-words vector space model (normal search). As such 800 images from the Malaysian on-line newspaper (under the Nation category) were indexed. On average, each document consists about 85 terms. Ten queries have been articulated, and for each query the relevant images is manually identified and labeled. The constructed queries resembled information related to 'person', 'location', 'event', 'time', 'organization' and 'position'. The query is in the form short natural language sentence such as "*Voter in the 12th general election*". The popular precision and recall measures were used.

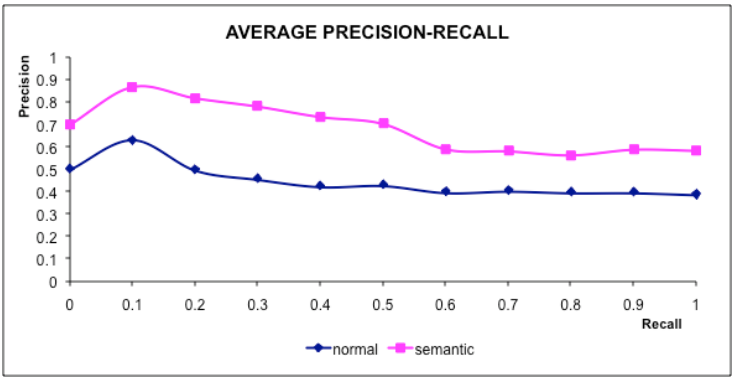


**Fig. 5.** Comparison of precision at standard recall values

Fig 5 illustrates the average precision for the 10 queries at a standard recall level. The result clearly shows that the semantic search approach significantly perform better than the normal approach, with an average precision of 7.50 and 4.90 respectively.

## 5 Discussion and Conclusion

Semantically annotated images based on the low level features such as color and texture are still difficult to perform and little success development has been reported. The more practical approach is to use the surrounding text for capturing the textual annotations of images and subsequently extends such annotations to lexical ontology or lexical resources. In this paper, we described an approach for extracting such textual meaning of digital images from the textual description and the surrounding text. The textual meaning is then semantically enhanced by mapping them to WordNet and common-sense concepts (or topics) that are derived from the ConceptNet lexical concepts. A set of NER patterns were designed in order to identify and differentiate important named entities in images. We scope our domain to the 'Nation' category of the Malaysian on-line newspapers. Evaluation on the initial 800 images has shown promising result, but further evaluation is definitely required with more queries, larger data sets and different domains.

WordNet has been the preferred lexical resource among information retrieval researchers for supporting semantic search and query expansion. However, not all terms can be directly associated with WordNet synsets. While WordNet can provide structured semantic meanings for some terms or concepts, images can go beyond such meanings to common-sense concepts or knowledge. Our approach has shown that how topics or common sense concepts can be derived from the ConceptNet topic_gisting module in order to enhance the semantic search.

It is interesting to see how multilingual or even cross language retrieval can be supported in semantic annotation [21]. Malaysian on-line newspapers can be grouped into two main languages the English (which is presented in this paper) and the Malay. Therefore, our current research work is to support cross language of semantic annotation for these languages.

**Acknowledgments.** We would like to thank the Malaysian's Ministry of Science, Technology and Innovation (MOSTI) and Institute of Higher Learning (IHL) for supporting this research project.

## References

1. Inoue, M.: On the need for annotation-based image retrieval. In: Workshop on Information Retrieval in Context (2004)
2. Benitez, A.B., Chang, S.-F.: Semantic Knowledge Construction from Annotated Image Collections. In: Proceedings of the 2002 International Conference on Multimedia & Expo (ICME 2002), Lausanne, Switzerland, August 26-29 (2002)
3. Varelas, G., Voutsakis, E., Raftopoulou, P., Petrakis, E.G.M., Evangelos, E., Milios, E.E.: Semantic similarity methods in WordNet and their application to information retrieval on the web. In: Proceedings of the 7th Annual ACM International Workshop on Web Information and Data Management, pp. 10–16 (2005)
4. Stokoe, C., Oakes, M.P., Tait, J.: Word Sense Disambiguation in Information Retrieval Revisited. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, Toronto, Canada, pp. 159–166 (2003)
5. Gong, Z., Cheang, C.W., Leong, H.U.: Web query expansion by WordNet. In: Andersen, K.V., Debenham, J., Wagner, R. (eds.) DEXA 2005. LNCS, vol. 3588, pp. 166–175. Springer, Heidelberg (2005)
6. Liu, H., Singh, P.: ConceptNet – A Practical Commonsense Reasoning Toolkit. BT Technology Journal 22(4), 211–226 (2004)
7. Hsu, M.-H., Tsai, M.-F., Chen, H.-H.: Query Expansion with ConceptNet and WordNet: An Intrinsic Comparison. In: Ng, H.T., Leong, M.-K., Kan, M.-Y., Ji, D. (eds.) AIRS 2006. LNCS, vol. 4182, pp. 1–13. Springer, Heidelberg (2006)
8. Eagle, N., Singh, P., Pentland, A.: Common sense conversations: understanding casual conversation using a common sense database. In: Proceedings of the Artificial Intelligence, Information Access, and Mobile Computing Workshop (2003)
9. Gong, Z., Leong, H.U., Cheang, C.W.: Web Image Indexing by Using Associated Data. Knowledge and Information Systems 10(2), 243–264 (2006)
10. Zhigang, H., Wang, X.-J., Lui, Q., Lu, H.: Semantic Knowledge Extraction and Annotation for Web Image. In: Proceedings of the 13th annual ACM international conference on Multimedia, Singapore, pp. 467–470 (2005)
11. Kahn, C.E., Rubin, D.L.: Automated Semantic Indexing of Figure Captions to Improve Radiology Image Retrieval. J. Am. Med. Inform. Assoc. 16, 380–386 (2009)
12. Edwardes, A.J., Purves, R.S.: Eliciting Concepts of Place for Text-based Image Retrieval. In: The 4th Workshop on Geographic Information Retrieval, pp. 15–17 (2007)
13. Deschacht, K., Moens, M.-F.: Text Analysis for Automatic Image Annotation. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pp. 1000–1007 (2007)
14. Feng, Y., Lapata, M.: Automatic Image Annotation Using Auxiliary Text Information. In: Proceedings of ACL 2008: HLT, pp. 272–280 (2008)
15. Gao, S., Wang, D.-H., Lee, C.-H.: Automatic Image Annotation through Multi-Topic Text Categorization. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, May 14-19, vol. 2, pp. II - II (2006)
16. Wang, C., Jing, F., Zhang, L., Zhang, H.-J.: Scalable search-based image annotation. Multimedia Systems 14, 205–220 (2008)

17. Hsu, M.-H., Chen, H.H.: Information Retrieval with Commonsense Knowledge. In: Proceedings of the SIGIR 2006, pp. 651–652 (2006)
18. Kuo, C.-H., Chou, T.-C., Tsao, N.-L., Lan, Y.-H.: CANFIND: A Semantic Image Indexing and Retrieval System. In: Proceedings of the 2003 International Symposium on Circuits and Systems, ISCAS 2003, vol. 2, pp. 644–647 (2003)
19. Kawata, K., Sakai, H., Masuyama, S.: QUARK: A Question and Answering System Using Newspaper Corpus as a Knowledge Source. In: Proceeding of the Third NTCIR Workshop (2003)
20. Liu, H., Lieberman, H.: Robust Photo Retrieval Using World Semantics. In: Proceedings of the LREC 2002 Workshop on Creating and Using Semantics for Information Retrieval and Filtering, Canary Islands (2002)
21. Sacaleanu, B., Buitelaar, P., Volk, M.: A Cross-Language Document Retrieval System Based on Semantic Annotation. In: Proceedings of the 10th conference on European Chapter of the Association for Computational Linguistics, pp. 231–234 (2003)