Bridging Semantic Gaps in Information Retrieval: Context-Based Approaches

Cam-Tu Nguyen
Graduate School of Information Science
Tohoku University, Japan
ncamtu@ecei.tohoku.ac.jp

supervised by Prof. Takeshi Tokuyama
Graduate School of Information Science
Tohoku University, Japan
tokuyama@dais.is.tohoku.ac.jp

ABSTRACT

In Information Retrieval (IR), the semantic gap is the difference between what computers store and what users expect via their queries. There are several reasons for the existence of those gaps such as homonymy and synonymy in text retrieval, or the typical difference between low-level representations and keyword-based queries in image retrieval. The objective of this work is to close these gaps by effective, scalable and not-so-expensive solutions. The main idea is to exploit available unstructured data and hidden topic models to infer surrounding contexts for better information retrieval (in both text retrieval and image retrieval). Early results obtained on two problems, namely Web search clustering and image annotation, show the effectiveness of the proposed approaches.

1. INTRODUCTION

The semantic gap characterizes the difference between two (or more) descriptions of an object in different representations [1]. This problem, which is underlined in many aspects of Web mining and Information Retrieval (IR), has posed a lot of challenges. In text retrieval, the gap is commonly caused by some natural linguistic phenomena such as synonymy or homonymy. Synonymy that is two or more different words have similar meanings causes difficulty in matching two related documents. For example, the similarity between two documents (particularly the short ones) containing "movie" and "film" is probably lower than what we expect. On the other hand, homonymy means a word can have multiple meaning. One example is the word "bank" exists in several contexts such as "organization" or "river side". Consequently, we may accidentally put an advertising message about a bank (an organization) on a Web page about bank (of some river).

The problem of semantic gap is more typical in Multimedia Mining and Image/Video Retrieval [6]. While early content-based image retrieval systems were based on the query-by-example schema, which formalizes the task as search for best matches to example images provided by users, the attention now moves to query-by-semantic schema in which queries are provided in natural language. This leads to the problem of automatic image annotation, which is concerned with assigning labels to images for later retrieval. Due to the semantic gap between low-level image representations (such as color, contour, shape) and high-level concepts (tiger, mountain, etc.), IR researchers commonly find image annotation a difficult problem to cope with.

This work is a thorough investigation of how surrounding contexts can help to bridge semantic gaps in Web Mining and Information Retrieval. Unlike previous works using knowledge bases or ontology, which are very helpful but not available in many domains and languages, this work focuses on not-so-expensive solutions by making use of unstructured-but-available data, which are easier to acquire, as well as recent advances of topic modeling methods. The targeted contributions consist of the following main points:

1. A careful review: summarizing problems and solutions related to semantic gaps and solutions in the literature regarding text and image retrieval.

2. Contextual matching toward text retrieval: providing a framework [24] for web search clustering and labeling based on Latent Dirichlet Allocation (LDA) [3]. The main idea is to perform topic analysis for a large and easy-to-obtain collection and use those topics to capture word relationships for better semantic matching. The advantages of this framework include (1) reducing the sparseness of Web search snippets; (2) reducing data mismatching caused by synonymy, homonymy, abbreviation words, etc.; (3) easy to use and adaptable to other languages and applications [25, 15].

3. Context-based image annotation toward image retrieval: introducing a feature-word-topic model for image annotation based on mixture hierarchies and topic modeling [23]. The model is able to guess the context from the scene for better annotation while keeping overhead computations reasonable. The separation of topics (of words) and low-level image representations makes the model more adaptable to other feature selection methods as well as topic analysis methods.

4. Context from auxiliary texts: developing the feature-word-topic model to capture richer relationships (hierarchy, correlated topics) and exploit multimodality in image annotation. Images in real-world are likely accompanied with metadata such as tags, filenames,

VLDB 2010 PhD Workshop, September 13, 2010, Singapore

which bring very useful evidences to infer the context of images. Our method differs from previous multimodal methods [8, 13, 17] is that we only consider a part of the whole vocabulary for annotation, which are subsequently referred as annotation vocabulary, but build a topic model on the whole vocabulary. By doing so, words, which are not in the annotation vocabulary but appear in tags or filenames of an image, can be exploited to find out the context of the image. Thanks to the topic modeling at word-level, the approach is able to limit the annotation vocabulary to only significant concepts while allowing to search images by synonyms or related words.

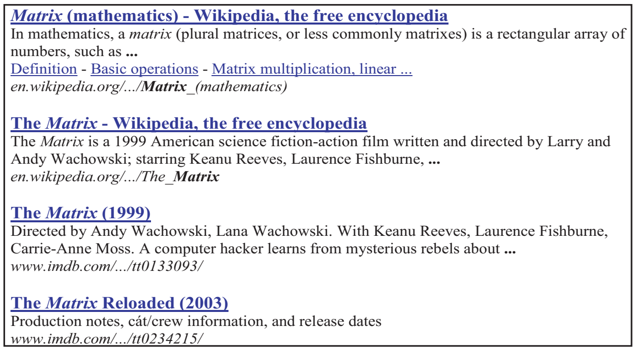
5. Context from global features: the importance of image features in contextual inference will be taken into consideration for image retrieval. Recent studies [7, 31] have showed the role of context detection via global features (such as GIST descriptors) for example-based image retrieval and object detection. Another objective of this work is to further study the connections of such types of global features with topics to improve image annotation.

The rest of this paper is divided into four parts. Section 2 will list some (but a few) related works on the attempts to bridge semantic gaps in text and image retrieval. Section 3 and 4 will present in more details what have been done up to now corresponding to parts 2 and 3 in the above list. Descriptions of on-going-work and future work will be given in Section 5.

2. RELATED WORKS

There have been a lot of attempts to close the gaps in Web Mining and Information Retrieval. Query expansion approach aims at improving the relevance of returned documents by expanding queries based on concept thesaurus, word co-occurrence statistics, query logs, and relevance feedback [20]. Latent semantic analysis (LSA) has been used to map words into the concept space so as to improve the relevance of retrieved results [16, 20]. Other studies use taxonomy, ontology and knowledge base to represent the semantic correlation between words for better contextual matching [2, 29, 9].

In the context of image retrieval, many noticeable approaches have been proposed to bridge the "semantic gap" for better image retrieval. Some approaches attempt to reduce annotation errors by making use of word relationships such as {fish, ocean} and {desert, sand}[18, 14, 38, 37, 36]. Other approaches make use of external resources such as auxiliary texts of web images [35, 26, 8], Wordnet and ontology [14, 26], Google distance [36], click through data [32], and Wikipedia articles [27]. Topic-based approaches model joint distributions of visual features and words [4, 21, 11, 12]. On the other hand, Multiple Instance Learning (MIL) approaches [5, 30] focus on solving the problem of "weakly labeling" in image annotation that is the lack of correspondence between labels and regions in images. Multiple feature types [31, 7, 19] are also selected to improve image retrieval, and object detection/ location. Recently, studies [34, 31] on jointly modeling scene classification and image annotation (or object detection) have been conducted on the attempt to exploit the context from the scene.



***Matrix* (mathematics) - Wikipedia, the free encyclopedia**
In mathematics, a *matrix* (plural matrices, or less commonly matrixes) is a rectangular array of numbers, such as **...**
Definition - Basic operations - Matrix multiplication, linear ...
*en.wikipedia.org/.../**Matrix**_(mathematics)*

The *Matrix* - Wikipedia, the free encyclopedia
The *Matrix* is a 1999 American science fiction-action film written and directed by Larry and Andy Wachowski; starring Keanu Reeves, Laurence Fishburne, **...**
*en.wikipedia.org/.../The_**Matrix***

The *Matrix* (1999)
Directed by Andy Wachowski, Lana Wachowski. With Keanu Reeves, Laurence Fishburne, Carrie-Anne Moss. A computer hacker learns from mysterious rebels about **...**
www.imdb.com/.../tt0133093/

The *Matrix* Reloaded (2003)
Production notes, cát/crew information, and release dates
www.imdb.com/.../tt0234215/

Figure 1: Search results for the query "matrix" from Google

The general idea of our approach, which differs from most of previous works, is that we exploit and develop topic models to capture various relationships among words (either textual or visual ones) for contextual indexing, matching in IR. Recently, topic modeling has been receiving significant interest in text mining and become a very useful tool to analyze word relationships. This research focuses on exploiting the contextual aspects of topics to bridge semantic gaps in IR-related problems. In comparison with Wordnet and other word-to-word resources, topics are easier to obtain and pay more attention to contextual information. For example, the relationships such as "grass" (background) and "tiger" (foreground object) in images, which are likely not available in predefined relationships of Wordnet, can be detected to reflect the common combinations of concepts to form scenes.

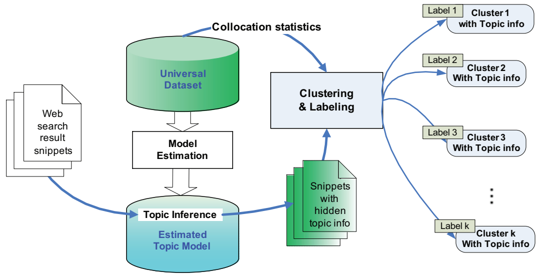
3. ENHANCING CONTEXTUAL MATCHING TOWARD TEXT RETRIEVAL

3.1 Motivation

Enhancing contextual matching plays an important role in ranking, and organizing Web pages in text retrieval. As a matter of fact, queries, search results (also called "snippets"), tags and captions of images, or advertisements (for content online advertisement) are very short texts, that is, they consist of from a dozen words to a few sentences. Consequently, they do not provide enough shared-context for good matching, ranking or organizing. Toward enhancing contextual matching, we focus on the problem of Web search clustering and labeling in order to demonstrate how contextual information can help to meet user requirements.

Web search clustering is a solution to reorganize search results (also called "snippets") in a more convenient way for browsing. There are three key requirements for such post-retrieval clustering systems: (1) The clustering algorithm should group similar documents together; (2) Clusters should be labeled with descriptive phrases; and (3) The clustering system should provide high quality clustering without downloading the whole Web pages.

These requirements and the third one in particular introduce several challenges to clustering. In contrast to normal documents, these snippets are usually noisier, less topic-focused, and much shorter. Figure 1 demonstrates four snippets returned from Google for the query "matrix". It can be observable that those snippets belong to two different clus-


Collocation statistics
Universal Dataset
Web search result snippets
Model Estimation
Topic Inference
Estimated Topic Model
Clustering & Labeling
Snippets with hidden topic info
Label 1
Cluster 1 with Topic info
Label 2
Cluster 2 With Topic info
Label 3
Cluster 3 With Topic info
Label k
Cluster k With Topic info

(a) *Choosing an appropriate "universal dataset"*
(b) *Performing topic analysis for the universal dataset*
(c) *Finding collocations in the universal dataset*
(d) *Performing topic inference for search snippets*
(e) *Combining the original snippets with their hidden topics*
(f) *Building a clustering/labeling system on the enriched snippets*

Figure 2: The general framework of clustering Web search results with hidden topics

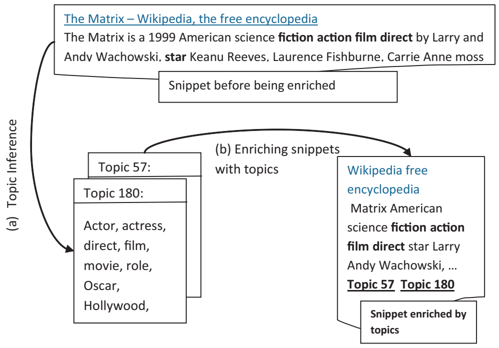

The Matrix – Wikipedia, the free encyclopedia
The Matrix is a 1999 American science fiction action film direct by Larry and Andy Wachowski. star Keanu Reeves. Laurence Fishburne. Carrie Anne moss
Snippet before being enriched
(a) Topic Inference
(b) Enriching snippets with topics
Topic 57:
Topic 180:
Actor, actress, direct, film, movie, role, Oscar, Hollywood,
Wikipedia free encyclopedia
Matrix American science fiction action film direct star Larry Andy Wachowski, …
Topic 57 Topic 180
Snippet enriched by topics

Figure 3: Example of enriching a snippet with hidden topics estimated from the Universal dataset

ters, one relates to film and the other is about mathematics. The first and second snippets are obtained from the same website (Wikipedia), thus share the title even that they are about different topics. On the other hand, although the forth snippet is about film but it does not have any common words with the other snippets about the same topic. These observations show that shallow matching based on words cannot obtain desirable results.

3.2 The Proposed Method

We present a framework which is based on recent successful topic analysis models, such as Probabilistic-Latent Semantic Analysis, or Latent Dirichlet Allocation. The underlying idea of the framework is that we collect a very large external data collection called "Universal Dataset", and then build a clustering and labeling system on both the original snippets and a rich set of hidden topics discovered from the universal data collection. The hidden topics from "Universal Dataset", which play as background knowledge, bring a richer representation of snippets to be clustered (see Figure 3). The framework is summarized in Figure 2, and consists of six major steps.

3.3 Discussion

The main advantages of the framework include the following points:

- Reducing data sparseness: different word choices make snippets of the same topic less similar, hidden topics do make them more related than the original.
- Reducing data mismatching: some snippets sharing unimportant words, which could not removed completely in the phase of stop word removal, are likely close in similarity. By taking hidden topics into account, the pairwise similarities among such snippets are decreased in comparison with other pairs of snippet. As a result, this goes beyond the limitation of shallow matching based on word/lexicon.
- Providing informative and meaningful labels: In this work, we use topic similarity between terms/ phrases and the cluster as an important feature to determine the most suitable label, thus providing more descriptive labels.
- Adaptable to another languages: The framework is simple to implement. All we need is to collect a large-scale data collection to serve as the universal data and exploit the topics discovered from that dataset as additional knowledge in order to measure similarity between snippets.
- Easy to reuse: The remarkable point of this framework is the hidden topic analysis of a large collection. This is totally unsupervised process but still takes time for estimation. However, once estimated, the topic model can be applied to more than one task which is not only clustering and labeling but also classification, contextual matching, etc.

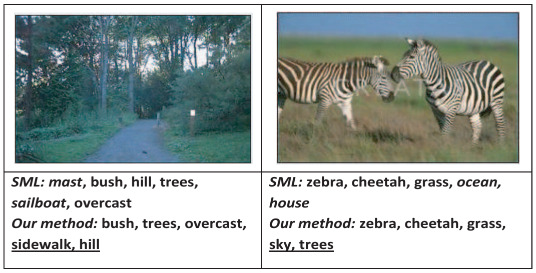
The proposed method for clustering and labeling in Vietnamese has been published in [24]. Other similar frameworks have also been used for short text classification (in English) and online advertisement (in Vietnamese) in our publications [25, 15], which showed the highly adaptability of the approach.

4. INFERRING CONTEXTS FROM SCENES FOR IMAGE ANNOTATION

4.1 Motivation

Image annotation is to automatically associate semantic labels with images in order to obtain a more convenient way for indexing and searching images on the Web. This has been an active topic of ongoing research during the last decade and led to several noticeable methods. Among the existing methods, Supervised Multiclass Labeling (SML) [5], which is based on mixture hierarchies, is considered as a state of art thanks to its scalability and direct optimality toward annotation. However, SML does not consider word relationships in annotation, where the associations such as {beach, sand, sun}, or {building, street, car, people} can be used to remove noisy annotations.

In order to take advantages of SML while exploiting word relationships for better annotations, we propose a method based on mixture hierarchies and probabilistic Latent Semantic Analysis (pLSA). Applying pLSA to words of images, we are able to estimate word-topic distributions and use them to reduce image annotation errors. Take the leftmost picture in Figure 4 as an example, branches are confused with masts (in SML method) due to visual similarities.



SML: mast, bush, hill, trees, *sailboat*, overcast
Our method: bush, trees, overcast, sidewalk, hill

SML: zebra, cheetah, grass, *ocean*, *house*
Our method: zebra, cheetah, grass, sky, trees

Figure 4: Examples of annotations of SML and our method

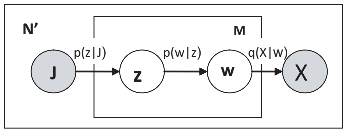


Figure 5: Feature-Word-Topic Model for Annotation in which N' is the number of new image J, and M is the number of annotation candidates selected using $p(w|X)$

Our method, however, infers "forest" as a dominant topic and lowers rankings of "mast" and "sailboat", thus provides better annotations.

4.2 Problem Statement and Notations

Given a vocabulary of words V, an image I is represented by a set of feature vectors X_I and can be annotated by a set of words W_I from V. $D = \{I_1, I_2, \ldots, I_N\}$ is a collection of annotated images. That means every I_n has manually assigned to a word set W_{I_n}. Suppose that we have a new image J represented by X_J, the problem of image annotation is to automatically assign a set of words W_J to J. Here, for simplicity, we refer to X_J and W_J as X and W.

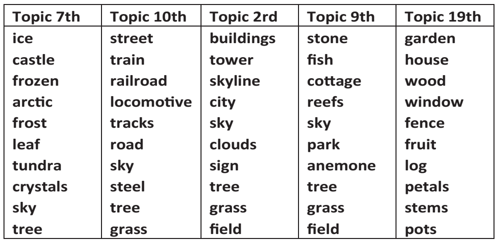
4.3 The Proposed Method

The training phase makes use of training images, which are images manually annotated with labels, to train two models: (1) a feature-word distribution $p(\mathbf{x}|w)$ using mixture hierarchies and MIL [5, 33]; (2) a word-topic distribution using labels of images in the training dataset and pLSA [10].

We then propose a novel model, see Figure 5, to annotate new images (in the annotation phase). In order to annotate image J, we first select top M annotation candidates using $p(X|w)$. The M words are then used to infer dominant topics of image J. Finally, we rerank M annotation candidates using both feature-word distribution and dominant topics of image J.

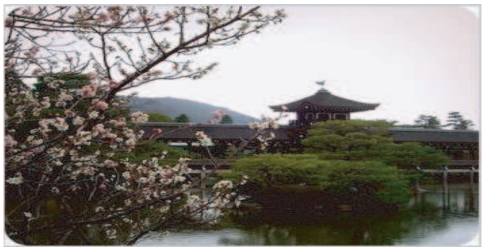
4.4 Discussion

In order to demonstrate our idea, let us consider the example given in Figure 6 and 7. Figure 6 shows some topics estimated from labels in the training dataset obtained from Flickr. In Figure 7, we show how those topics can be used to improve annotation performance. Based on feature-word distributions, the top 20 words are selected and shown in the figure. We can see that the visual representation gives wrong interpretation of the picture, which results in words like "ruin", "town" at first and second positions. Other reasonable interpretation of the picture makes the topic describing this scene (topic 23) surpass the other topics. By taking topics into account, the more reasonable words ("temple", "pagoda") have higher ranking positions than only based on features. Due to the "semantic gap", the visual representation is not good enough for image annotation. We need more "context" from scene to infer reasonable labels. More details about this method are given in our submitted paper[23], which is available on our website[1].



Topic 7th	Topic 10th	Topic 2rd	Topic 9th	Topic 19th
ice	street	buildings	stone	garden
castle	train	tower	fish	house
frozen	railroad	skyline	cottage	wood
arctic	locomotive	city	reefs	window
frost	tracks	sky	sky	fence
leaf	road	clouds	park	fruit
tundra	sky	sign	anemone	log
crystals	steel	tree	tree	petals
sky	tree	grass	grass	stems
tree	grass	field	field	pots

Figure 6: Topics estimated from the training dataset in Flicker dataset



(a) Top 20 words from feature-word distributions:
ruins; town; sculpture; castle; Buddhist; temple; house; baby; pagoda; fruit; autumn; woman; formula; street; beach; umbrella; fountain; fence; Buddha; indian
(b) Top 5 words after refinement with topics
temple; Buddhist; Buddha; town; pagoda
(c) Top Topics
Topic 23rd: temple, buddhist, budda, monks, figures, pagoda, tree
Topic 13th: town, door, umbrella, stair, monastery, window,
Topic 45th: statue, sculpture, plant, car, fountain, prob, tree, glass
Topic 46th: park, ruins, face, sky, tusk, tree, field, grass, cloud
(d) Manual Annotation: temple, garden, shrine, palace

Figure 7: Demonstration of how topics can be used to improve annotation performance

In recent years, there are a lot of applications of topic models, which originated from text mining, in image-related problems. Most of the current approaches model directly topic-feature distributions [4, 11, 12, 21, 22, 17]. If continuous features are used [4, 12], topic estimation becomes very complicated and expensive (in terms of time complexity) since the feature space is very large in comparison with word space. If features are clustered to form discrete visual-words [11, 17, 21], the clustering step on a large dataset of images is also very expensive. Moreover, topics of features are usually hard to interpret than topics of words.

The difference of our method from previous approaches is that we model topics via words, not features (see Figure 8). As a result, we do not need to modify topic models for training, where captions are available. To infer topics for an unannotated image, we only need to consider $p(\mathbf{x}|w)$ as weight for w instead of word occurrence in the original mod-

[1] http://www.dais.is.tohoku.ac.jp/˜ncamtu/research.htm

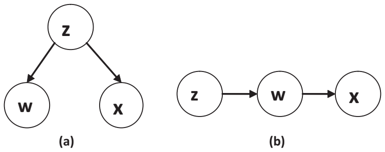


Figure 8: The difference of our method in comparison with other topic-based approaches to image annotation: (a) Other approaches; (b) Our method

els. Since feature-word distribution for a word is estimated based on a subset of the training dataset containing that word, it is more practical in comparison with visual-word construction.

Since all we need from the low-level feature representation is the feature-word distribution $p(x|w)$, other methods such as [28] can be used to obtain this purpose.

5. CONCLUSION AND FUTURE WORKS

Semantic of an object depends on the context it is regarded within [1]. This work is an investigation of multiple methods to exploit context for information retrieval (in both text and image retrieval). To complete the story, further studies need to be conducted according to the objectives described in Section 1.

On the next attempt, the feature-word-topic model will be modified to benefit from auxiliary texts and richer relationships (such as hierarchy, correlated) between topics. In other words, we will perform more complicated topic modeling on a larger vocabulary, which includes both annotation words and surrounding texts (or name of image files). Since we consider only M selected candidates, which are in the annotation vocabulary, the annotation step works exactly the same as described. Due to computation complexity and the dynamic of human language, the annotation vocabulary is usually limited. By modeling topics for a larger vocabulary, we are able to infer topics based on surrounding texts and features (via feature-word distributions). In fact, the surrounding text may be not enough for searching but can be used as a hint for annotation refined by topics. For example, suppose that we model topics with an extended vocabulary containing “Eiffel”, an image file name “Eiffel” should increase the distributions for topics related to “tower”, “city” even if “Eiffel” is not in the annotation vocabulary. This property also allows us to search with queries that are not in the annotation.

Finally, we will extend feature-word-topic model to include many types of feature selection, especially global features [31, 34] to improve the quality of scene detection. One feature representation can be considered as one view of an image, different views of an image can be used to obtain better annotation. For example, we can train one model of $p(\mathbf{x}|w)$ for local feature descriptors (such as SIFT, DCT, and so on), one model $q(\mathbf{y}|w)$ for global features (such as contour, shapes, Gist). Weighted candidates from different views can be selected, merged and refined for annotation using topics. Considering an image in different views not only help to improve annotation performance but also reduce the time complexity to estimate feature-word distribution. Instead of using feature vectors with large dimension, we can divide them to several types of feature vectors, each of which has smaller dimension.

6. REFERENCES

[1] http://en.wikipedia.org/wiki/Semantic_gap, 2010.

[2] S. Banerjee, K. Ramanathan, and A. Gupta. Clustering short texts using wikipedia. In *SIGIR ’07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 787–788, New York, NY, USA, 2007. ACM.

[3] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *J. Machine Learning Research*, 3:993–1022, 2003.

[4] D. M. Blei and M. I. Jordan. Modeling annotated data. In *SIGIR ’03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 127–134, 2003.

[5] G. Carneiro, A. B. Chan, P. J. Moreno, and N. Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(3):394–410, 2007.

[6] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv.*, 40(2):1–60, 2008.

[7] M. Douze, H. Jégou, H. Sandhawalia, L. Amsaleg, and C. Schmid. Evaluation of gist descriptors for web-scale image search. In *CIVR ’09: Proceeding of the ACM International Conference on Image and Video Retrieval*, pages 1–8, New York, NY, USA, 2009. ACM.

[8] Y. Feng and M. Lapata. Automatic image annotation using auxiliary text information. In *Proceedings of ACL-08: HLT*, pages 272–280, Columbus, Ohio, June 2008. Association for Computational Linguistics.

[9] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJCAI’07: Proceedings of the 20th international joint conference on Artifical intelligence*, pages 1606–1611, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc.

[10] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1-2):177–196, 2001.

[11] E. Hörster, R. Lienhart, and M. Slaney. Image retrieval on large-scale image databases. In *CIVR ’07: Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 17–24, New York, NY, USA, 2007. ACM.

[12] E. Hörster, R. Lienhart, and M. Slaney. Continuous visual vocabulary modelsfor plsa-based scene recognition. In *CIVR ’08: Proceedings of the 2008 international conference on Content-based image and video retrieval*, pages 319–328, New York, NY, USA, 2008. ACM.

[13] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation of news images with large vocabularies and low quality training data. In *Proceedings of ACM Multimedia*, 2004.

[14] Y. Jin, L. Khan, L. Wang, and M. Awad. Image annotations by combining multiple evidence & wordnet. In *MULTIMEDIA ’05: Proceedings of the*

13th annual ACM international conference on Multimedia, pages 706–715, New York, NY, USA, 2005. ACM.

[15] D.-T. Le, C.-T. Nguyen, Q.-T. Ha, X.-H. Phan, and S. Horiguchi. Matching and ranking with hidden topics towards online contextual advertising. In *WI-IAT '08: Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, pages 888–891, Washington, DC, USA, 2008. IEEE Computer Society.

[16] T. A. Letsche and M. W. Berry. Large-scale information retrieval with latent semantic indexing. *Inf. Sci.*, 100(1-4):105–137, 1997.

[17] R. Lienhart, S. Romberg, and E. Hörster. Multilayer plsa for multimodal image retrieval. In *CIVR '09: Proceeding of the ACM International Conference on Image and Video Retrieval*, pages 1–8, New York, NY, USA, 2009. ACM.

[18] J. Liu, B. Wang, H. Lu, and S. Ma. A graph-based image annotation framework. *Pattern Recogn. Lett.*, 29(4):407–415, 2008.

[19] A. Makadia, V. Pavlovic, and S. Kumar. A new baseline for image annotation. In *ECCV '08: Proceedings of the 10th European Conference on Computer Vision*, pages 316–329. Springer-Verlag, 2008.

[20] C. D. Manning, P. Raghavan, and H. Schutze. *An Introduction to Information Retrieval.* Cambridge University Press, 2009.

[21] F. Monay and D. Gatica-Perez. Plsa-based image auto-annotation: constraining the latent space. In *MULTIMEDIA '04: Proceedings of the 12th annual ACM international conference on Multimedia*, pages 348–351, New York, NY, USA, 2004. ACM.

[22] F. Monay and D. Gatica-Perez. Modeling semantic aspects for cross-media image indexing. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(10):1802–1817, 2007.

[23] C.-T. Nguyen, N. Kaothanthong, X.-H. Phan, and T. Tokuyama. A feature-word-topic model for image annotation. submitted, July 2010.

[24] C.-T. Nguyen, X.-H. Phan, S. Horiguchi, T.-T. Nguyen, and Q.-T. Ha. Web search clustering and labeling with hidden topics. *ACM Transactions on Asian Language Information Processing (TALIP)*, 8(3):1–40, 2009.

[25] X.-H. Phan, C.-T. Nguyen, D.-T. Le, L.-M. Nguyen, S. Horiguchi, and Q.-T. Ha. A hidden topic-based framework towards building applications with short web documents. *IEEE Transactions on Knowledge and Data Engineering*, 99(PrePrints), 2010.

[26] A. Popescu, C. Millet, and P.-A. Moëllic. Ontology driven content based image retrieval. In *CIVR '07: Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 387–394, New York, NY, USA, 2007. ACM.

[27] T. Quack, B. Leibe, and L. Van Gool. World-scale mining of objects and events from community photo collections. In *CIVR '08: Proceedings of the 2008 international conference on Content-based image and video retrieval*, pages 47–56, New York, NY, USA, 2008. ACM.

[28] V. C. Raykar, B. Krishnapuram, J. Bi, M. Dundar, and R. B. Rao. Bayesian multiple instance learning: automatic feature selection and inductive transfer. In *in Proceedings of the 25th International Conference on Machine learning*, pages 808–815. ACM, 2008.

[29] P. Schonhofen. Identifying document topics using the wikipedia category network. In *WI '06: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 456–462, Washington, DC, USA, 2006. IEEE Computer Society.

[30] V. Stathopoulos and J. M. Jose. Bayesian mixture hierarchies for automatic image annotation. In *ECIR '09: Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*, pages 138–149, Berlin, Heidelberg, 2009. Springer-Verlag.

[31] A. Torralba, K. P. Murphy, and W. T. Freeman. Using the forest to see the trees: exploiting context for visual object detection and localization. *Commun. ACM*, 53(3):107–114, 2010.

[32] T. Tsikrika, C. Diou, A. P. de Vries, and A. Delopoulos. Image annotation using clickthrough data. In *CIVR '09: Proceeding of the ACM International Conference on Image and Video Retrieval*, pages 1–8, New York, NY, USA, 2009. ACM.

[33] N. Vasconselos. Image indexing with mixture hierarchies. In *IEEE Conference in Computer Vision and Pattern Recognition*, 2001.

[34] C. Wang, D. Blei, and F.-F. Li. Simultaneous image classification and annotation. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 0:1903–1910, 2009.

[35] C. Wang, F. Jing, L. Zhang, and H.-J. Zhang. Image annotation refinement using random walk with restarts. In *MULTIMEDIA '06: Proceedings of the 14th annual ACM international conference on Multimedia*, pages 647–650, New York, NY, USA, 2006. ACM.

[36] Y. Wang and S. Gong. Refining image annotation using contextual relations between words. In *CIVR '07: Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 425–432, New York, NY, USA, 2007. ACM.

[37] Q. H. Zhiwu Lu, Horace Ip. Context-based multi-label image annotation. In *ACM International Conference on Image and Video Retrieval*, 2009.

[38] X. Zhou, M. Wang, Q. Zhang, J. Zhang, and B. Shi. Automatic image annotation by an iterative approach: incorporating keyword correlations and region matching. In *CIVR '07: Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 25–32, New York, NY, USA, 2007. ACM.