Automatic Generation of Semantic Fields for Annotating Web Images

Gang Wang[§♀], Tat Seng Chua[#], Chong-Wah Ngo[⊙], Yong Cheng Wang[♀]
♀Shang Hai Jiao Tong University
School of Computing, National University of Singapore
⊙Dept of Computer Science, City University of HongKong
§ Na Xun Hi-Tech Application Institute
wanggang_sh@hotmail.com,chuats@comp.nus.edu.sg,
cwngo@cs.cityu.edu.hk,ycwang@mail.sjtu.edu.cn

Abstract

The overwhelming amounts of multimedia contents have triggered the need for automatically detecting the semantic concepts within the media contents. With the development of photo sharing websites such as Flickr, we are able to obtain millions of images with user-supplied tags. However, user tags tend to be noisy, ambiguous and incomplete. In order to improve the quality of tags to annotate web images, we propose an approach to build Semantic Fields for annotating the web images. The main idea is that the images are more likely to be relevant to a given concept, if several tags to the image belong to the same Semantic Field as the target concept. Semantic Fields are determined by a set of highly semantically associated terms with high tag co-occurrences in the image corpus and in different corpora and lexica such as WordNet and Wikipedia. We conduct experiments on the NUS-WIDE web image corpus and demonstrate superior performance on image annotation as compared to the state-of-the-art approaches.

1 Introduction

The advancement in computer processor, storage and the growing availability of low-cost multimedia recording devices has led to an explosive growth of multimedia data. In order to effectively utilize such a huge amount of multimedia contents, we need provide tools to facilitate their management and retrieval. One of the most important tools is the automatic media concept detectors, which aim to assign high-level semantic concepts such as "bear" to the multimedia data. More formally, the concept detection for an web image is defined as: given a set of predefined concepts $\vec{C}$: [C_1, C_2 ...C_n], we assign a semantic concept C_i to the image if it appears visually in the image. Traditionally, such concept detectors are built by the classifier approaches. The performance of such detectors depends highly on the quality of training data. However, preparing a set of high quality training data usually needs a large amount of human labors. On the other hand, the social web is changing the way people create and use information. For example, users started to develop novel strategies to annotate the massive amount of multimedia information from the web. In image annotation, Kennedy et al. (2006) explored the trade-offs in acquiring training data by automated web image search as opposed to manual human labeling. Although the performance of systems with training data obtained by manual human labeling is still better than those whose training data is acquired by automated web search, the latter approaches have attracted many researchers' interest due to their potential in reducing human label efforts. However, the tags in the web images are known to be ambiguous and overly personalized (Matusiak 2006).

Figure 1 gives four examples to illustrate the relationships between the visual concept "bear" and the annotation tag "bear". Generally speaking, there are four types of relationships:

Coling 2010: Poster Volume, pages 1301–1309,
Beijing, August 2010

- The relevant tag: The user-tag "bear" properly reflects the content of an image, as shown in Figure 1(a). While "bear" has multiple senses, the visual concept corresponds directly to the most common sense of "bear".
- The ambiguous tag: The user-tag "bear" is ambiguously related to the visual content, as shown in Figure 1(b). In this example, the visual content is related to another sense of "bear": "a surly, uncouth, burly, or shambling person" (Merriam-Webster dictionary, 2010).
- The noisy tag: The user-tag "bear" is a noisy tag, as shown in Figure 1(c). In this example, the visual content is irrelevant to the concept "bear".
- The incomplete tag: The user-tag "bear" doesn't occur in the tag list of Figure 1(d). However, many human annotators believe that the visual concept "bear" exist in the Figure 1(d). Also, in Wikipeida, a panda is defined as a kind of a bear.

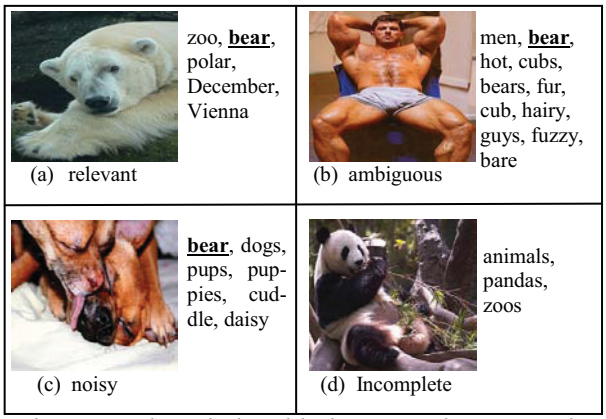


Figure 1: The relationship between the tags and the visual concept "bear" in NUS-Wide corpus.

In this paper, we aim to assign relevant tags to images in order to reduce the effects of ambiguous, noisy and incomplete tags. To distinguish relevant tags from other sense of tags, a common practice is to perform word sense disambiguation (WSD) to predict the right sense of a tag. Nevertheless, performing a WSD on a noisy and sparse set of tags, where the order and position of tags do not matter, is by no means easy. Most existing works on WSD, such as Navigli (2009) are based on clean data and word neighborhood statistics. They cannot be directly applied to address this problem. Although there are some works such as Wang et al. (2003) on capturing the semantics of noisy data, the problem of ambiguous words has not been considered. In addition, some semantic models such as PLSA (Hofmann 1999), LDA (David et al. 2003) have been proposed to capture the semantics. However, one challenge of employing such models is that there are many noisy tags in the web image domain. The reason for noisy tags is that the purpose of tagging is not only for content description, but also for other factors such as getting attention and so on (Ame and Naaman, 2007, Bischoff et al. 2008).

Given a web image with a tag list, we propose an approach to predict the "Semantic Field" of the image. Semantic Field (Jurafsky and Martin 2000) is designed to capture a more integrated relationship among the entire sets of tags. In our work, we consider four different cases of examples, as shown in Figure 1. In 1(a), the concept "bear" will be assigned to the image with relatively high probability, because "zoo", "bear", and "polar" provide clues that "bear" is the major focus of the image. In 1(b), the concept "bear" could possibly be disambiguated as not related to "animal", the most common sense of "bear", by investigating other tags such as "men","guys". In 1(c), the image will not be labeled as "bear", since the surrounding tags such as "dogs", "pups" do not support the existence of "bear" in the image. In 1(d), although the concept "bear" is missing, the image will be still labeled as "bear" since the surrounding tags such as "pandas", "animals", and "zoos" jointly suggest that "bear" appears in the image. The significance of user tags towards a target concept can be modeled from three different sources: the statistics from the web image corpus, Wordnet and Wikipedia. In summary, instead of directly matching the keywords and tags, we consider tags of an image collectively to predict the underlying semantic field. Ideally, the semantic field can highlight the major visual concepts in images so that we can assign the correct semantic labels to the images.

In the rest of this paper, we discuss related work in Section 2, while Section 3 reports the building of Semantic Fields and its application to web image ranking. Section 4 discusses the experimental setup and results. Finally, Section 5 contains our concluding remarks.

2 Related Work

In this section, we report the works on Semantic Field theory, text analysis in multimedia and the existing systems for a web image corpus.

2.1 Semantic Fields

Semantic Fields have been hotly debated in linguistics community (Grandy 1992, Garret 1992). Compared to lexical analysis, it considers the entire sets of words instead of a single word. The FrameNet project (Baker et al. 1998) is an attempt to realize the Semantic Field. However, the problem with FrameNet project is that it needs extensive human efforts to define the thematic roles for each domain and each frame, and hence it is domain specific.

2.2 Text Analysis in Multimedia

In multimedia, one of the important tasks is concept detection, which attempts to find the visual appearance of a concept such as "bear" in an image. However, due to the large variations in the low level visual feature space such as color, texture etc, in many cases, researchers are hardly able to capture the concept by visual information alone. Some researchers attempted to employ natural language analysis to detect the visual concept. Rowe (1994) explored the syntax of images' captions to infer the visual concepts present in images. For example, he found that the primary subject noun phrase usually denotes the most significant information in the media datum or its "focus". He assumed that both visual and text features will describe the same focus of the content. Wang et al. (2008) employed the similar idea to infer visual concepts in news video. They first aligned text information with visual information, and then captured the text focus to infer the visual concept. These works suggest that we can transfer the problem of visual concept detection to that of finding a text focus.

In addition, researchers proposed statistical models to combine text and visual features, such as the translation model (Duygulu et al. 2002, Jin et al. 2005), cross media relevance model (Jeon et al. 2003) and continuous relevance model (Lavrenko and Jeon, 2003). However, no matter what models we used, the annotation accuracy is still quite low, partially because of the existence of noise in tags. Jin et al. (2005) provided a solution to tackle such a noisy tag problem. They first investigated various semantic similarity measures between each keyword pairs in the tag list based on Wordnet. They then regarded non-correlated keywords as noises and discarded them. In this paper, there are three major differences between our work and the above work. First, because tags from Internet are not always included in Wordnet, we employ multi-resources of information to analyze the semantics. Second, we extend the analysis of the word pair relationship to the Semantic Field analysis. Third, since it is not easy to identify the noise in the tag list directly, we only analyze the tags which are highly relevant to the concept with a specific sense.

2.3 The State of the Art Systems

NUS-WIDE (Chua et al. 2009) is a large scale Web image corpus. It provides not only social tags from the web, but also the "gold" labels (or ground truth) for 81 concepts from large human labeling efforts. As far as we know, there are two reported systems that used the whole NUS-WIDE corpus to test their proposed methods. In Chua et al. (2009), the 81 concepts are detected by k nearest neighbor using the visual features of: color moments, color auto-correlogram, color histogram, edge direction histogram, wavelet texture, and a bag of visual words. The mean average precision (MAP) for the 81 concepts reaches 0.1569. Gao et al. (2009) extended the k-NN approach to use both text tags and visual information. For the tag information, they made use of the co-occurrence information to compute the probability of an image belonging to a contain concept. They used the same visual features as in (Chua et al. 2009). In their work, the taxonomy in WordNet is exploited to identify whether a target concept is generic or specific. The co-occurrence tag analysis is employed for generic concepts, while visual analysis is used for specific concepts. The MAP for this approach reaches 0.2887.

3 Building Semantic Fields for Annotating Web Images

In this paper, we attempt to capture text semantics collectively from the tag list of images to annotate their visual contents. Semantic Fields consist of a selected subset of the tag list and

the choice of these tags is based on their relevance to the contents of the targeted image with a specific sense. There are three characteristics in our Semantic Field model. First, the Semantic Field is built by only a subset of tag list. For example, the Semantic Field in Figure 1(a) is {zoo, bear, polar}. It could partially reduce the effect of the noise. Second, because inferring the visual concept of an image is more reliable by joint analysis of tags in the Semantic Field, rather than investigating one tag at a time in the whole tag list, we analyze the whole Semantic Field as a unit. By utilizing the context information in Semantic Field, the problems of ambiguous, noisy and incomplete tags are partially tackled. Third, we perform normalization to estimate the importance of Semantic Field, which is discussed in Section 3.1. If the value is large, it suggests that most of the tags in the image support the Semantic Field; that is, the probability that the target concept is the focus of the image is high, and vice versa. Such a design aims to minimize the effects of noisy and ambiguous tags.

3.1 A Probabilistic Model

We denote C_x as a target concept that appears in the content of an image. We want to determine the set of tags that are related to C_x from the user-supplied tags by building a Semantic Field SF_i for each image. The probability of the appearance of concept $P(C_x \mid SF_i)$ can be computed as:

$$P(C_x \mid SF_i) = \frac{P(SF_i \mid C_x) \times P(C_x)}{P(SF_i)} \qquad (1)$$

For the purpose of collecting and annotating images and simplifying the model, we did not consider the prior knowledge for each image. Thus, the prior probability $P(C_x)$ can be viewed as a constant with respect to a concept C_x. In addition, the range of the normalization factor $P(SF_i)$ is expected to be small, which will not affect the annotation of web images. This assumption is reasonable due to the fact that there are a large number of different tags, and these tags can be combined to form any Semantic Field in an arbitrary manner. The number of combinations is exponential to the number of possible tags available. This is also evident by the observation that most tag lists associated with the images are unique. In other words, two images with the same Semantic Field are seldom found in reality. With these in mind, Equation (1) can be approximated and simplified to:

$$P(C_x \mid SF_i) \propto P(SF_i \mid C_x) \qquad (2)$$

Given a Semantic Field SF_i, it may include n related tags $T_1, T_2, T_3, ..., T_n$. Thus Equation (2) is expanded to:

$$P(SF_i \mid C_x) = P(T_1, T_2, ...T_n \mid C_x) \qquad (3)$$

Two obvious approaches to compute Equation (3) are using the product of the individual terms or chain rule decomposition. However, we consider the individual terms to be interdependent and the chain rule decomposition is not easy to compute. To simplify the model, we employ the normalized linear fusion to expand Equation (3) as follows:

$$\mathrm{P}(T_1, T_2, ...T_n \mid C_x) = \frac{\sum_{i=1}^{n} P(T_i \mid C_x)}{\mathrm{TN}} \qquad (4)$$

The normalization factor is the total number (TN) of tags in the image tag list.

3.2 Using Multiple External Sources

To estimate the probability of a tag T_i given a target concept C_x, i.e., $P(T_i|C_x)$, we consider both the domain knowledge and general knowledge acquired from Internet. For the former, we utilize the co-occurrence statistics of tags in images which can be computed offline from any web image corpus. For the latter, we employ WordNet and Wikipedia for inferring the relatedness between tags and a target concept. Combining different knowledge sources, the probability is estimated as:

$$P(T_i \mid C_x) = P(T_{i\_wd} \mid C_x) \times P(T_{i\_wiki} \mid C_x) \times P(T_{i\_co} \mid C_x) \quad (5)$$

where T_{i_wd}, T_{i_wiki}, T_{i_co} represent the tag occurrences in WordNet, Wikipedia and co-occurrence statistics, respectively.

To compute Equation (5), we query different information sources using the target concept C_x. In WordNet, because the sense of the concept usually refers to the most common sense in our corpus, we choose the most common sense (noun) as the target. Using Figure 2 as an example, the concept "bear" is defined in WordNet as “massive plantigrade carnivorous or omnivorous mammals with long shaggy coats and strong claws”. In Wikipedia, with Figure 3 as an example, the related page is downloaded to

describe the concept "bear". For the co-occurrence statistics of the tag lists, we estimate their values from co-occurrence information from the image corpus. With the above knowledge, we compute the conditional probability of a tag being related to C_x as:

$$P(T_j \mid C_x) = \frac{\#(T_j, C_x)}{\#(C_x)} \qquad (6)$$

where j = {wd, wiki, co}, $\#(T_j, C_x)$ indicates the number of times the tag and the concept co-occur in an information source, and $\#(C_x)$ denotes the number of times the concept C_x appear in the information source. In addition, we employ an add-one smoothing approach [Jurafsky and Martin 2000] to further process the results.

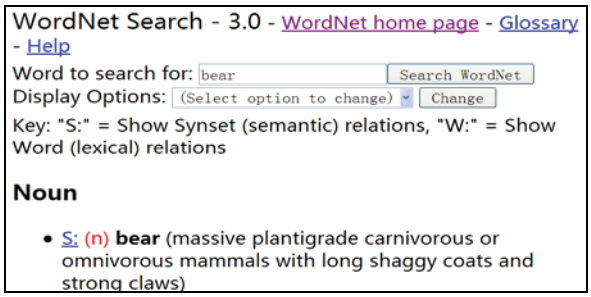


Figure 2: The information in WordNet

Figure 3: The information in Wikipedia.

Given a concept with a special sense, for all the tags in the corpus, we can obtain the conditional probabilities of each tag T_i based on Equation (5). We rank the tags according to $P(T_i \mid C_x)$. To reduce computations, we select the top N (N=200) tags as the highly related tags to a given concept and place them in a dictionary.

3.3 Building Semantic Field for Image Annotation

We now build the Semantic Fields to rank the images with respect to concept C_x. The detailed algorithm is shown in Figure 4.

Input:
1) Given a target concept, we rank all the tags in the corpus based on Equation (5).
2) Given a web image, we have a list of annotation tags ($l_1, l_2, ...l_{n1}$).

Step 1: Generate a dictionary (D) based on top N tags

Step 2: For (i=1; i<n1; i++)
If ($l_i \in D$) then put l_i into the Semantic Field for the image.

Step 3: Annotate the images and compute the probability of the occurrence of the concept via Equation (4)

Figure 4: The algorithm for building the Semantic Fields and annotating the images.

The algorithm comprises three steps:

1. bear	2. bears	3. polar	4. species
5. panda	6. cubs	7. giant	8. grizzly
9.teddy	10. pandas	…	…

Table 1: The top 10 tags for concept "bear" in most common sense.

First, given a target concept with a specific sense, we generate a dictionary based on the top N candidate tags as discussed in Section 3.2. Table 1 shows the top 10 tags in the dictionary for the concept "bear" with the most common sense. As we want to distinguish single and plural noun for different visual concepts, we do not employ the stemming algorithm. Although the results are not ideal, we find that many highly related words are included in the dictionary.

Second, we infer the annotation tags of the image from the dictionary and use that to build the Semantic Fields. Figure 1 demonstrates the resulting of Semantic Fields for images in Table 2.

Third, we assign the tags to images based on their Semantic Fields. Because most of the tags in Figure 1(a) and 1(d) are highly relevant to "bear" with the most common sense, we assign the semantics to these two images with high probabilities. Thus, the problem of incomplete tags is tackled in this case. On the other hand, since most of the tags in Figure 1(b) and 1(c) fail to support the concept "bear" with the most

common sense (the Semantic Field obtains less than 20% of tags' support), we only assign the semantics with very low probabilities. Thus, the ambiguous and noisy problem can be partially tackled.

Semantic Field for Figure 1 (a)	{zoo, bear, polar}
Semantic Field for Figure 1 (b)	{bear, bears}
Semantic Field for Figure 1 (c)	{bear}
Semantic Field for Figure 1 (d)	{animals, pandas, zoos}

Table 2: Semantic Fields of images in Figure 1.

4 Experiments

In this section, we first introduce the test-bed and measurement of the experiments. We then report the results and compare them with the state-of-the-art systems tested on NUS-WIDE corpus.

The NUS-Wide corpus (Chua et al. 2009) includes 269,648 images with 5,018 user-provided tags, and the ground-truth for 81 concepts for the entire database. These concepts are grouped into six different categories: graph, program, scene /location, event/activities, people and object. The choice of concepts is based on the generality and popularity in Flickr, the distributions in different categories and the common interests of the multimedia community. This corpus includes two parts. The first part contains 161,789 images to be used for training and the second part contains 107,859 images is used for testing.

The performance of the system is measured using the mean average precision (MAP) based on all the test images for all the 81 concepts. This is the same as the evaluation used in TRECVID. The MAP combines precision and recall into one performance value. Let $p^k = \{i_1, i_2, ..., i_k\}$ be a ranked version of the resulting set A. At any given rank k, let $R \bigcap p^k$ be the number of relevant images in the top k of p, where |R| is the total number of relevant images. Then MAP for the 81 concepts C_i is defined as:

$$MAP = \frac{1}{81}\sum_{C_i=1}^{81}[\frac{1}{|R|}\sum_{k=1}^{A}\frac{R \bigcap p^k}{k}\varphi(i_k)] \quad (6)$$

where the indictor function $\varphi(i_k) = 1$ if $i_k \in R$ and 0 otherwise.

4.1 Comparison with the State-of-the-Art Systems

We compare our approach against the reported systems on NUS-WIDE corpus.

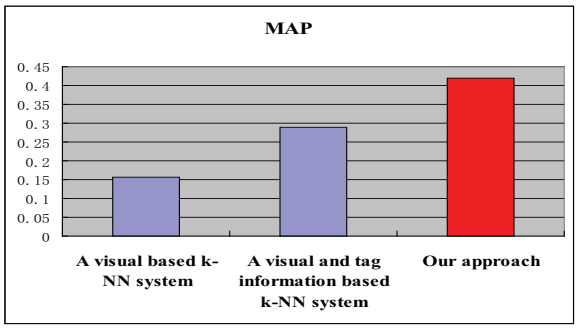
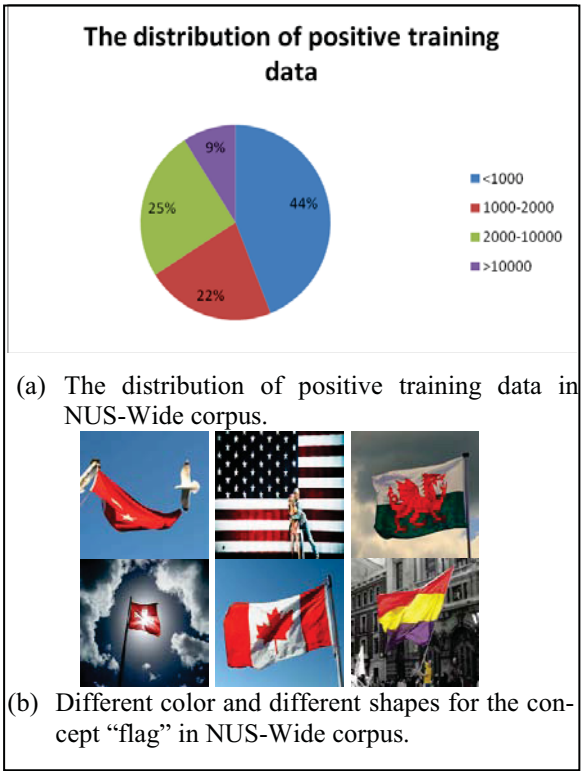


Figure 5: The comparison with the state-of-the-art system

In our approach, we employ the Semantic Field to annotate the images, which requires neither training data nor visual analysis, and is running directly on the test data. In contrast to the two previous approaches in Section 2.3, the input to Semantic Field is simply the tag list of an image. Figure 5 shows the performance comparisons among the three tested approaches. As compared to (Chua et al. 2009) and (Gao et al. 2009), which exhibit the best performance on NUS-WIDE so far, Semantic Field achieves a MAP of 0.4198 which shows a 45.4% improvement.

The reason for the superior performance of our approach is that there is insufficient training data, which means that most learning-based systems could not perform well. As seen in Figure 6(a), 44% of concepts have less than 1,000 positive training data. This is insufficient for training the classifiers for the visual concepts. Take the visual concept "flag" as the example. Considering that there are at least 200 national flags from different countries and regions, not to mention other types of flags such as holiday flag, there are large variations in concept "flag" as shown in Figure 6(b). Hence it is difficult to train a classifier with visual analysis by having only 214 positive training samples. This suggests that there may be a large

gap between the training and test data. On the other hand, because web images include not only visual features but also text information, we could employ text analysis to infer the visual concept. The advantages of our Semantic Field approach are that we could analyze multiple information sources to reduce the text variations and the performance of our approach is independent of the training data and visual features. With the increasing size of the corpus, the problems of few positive training data and large visual diversity between training and test data will be exacerbated. This is the reason why our approach is more robust than those based on visual analysis and traditional learning-based approaches.



(a) The distribution of positive training data in NUS-Wide corpus.

(b) Different color and different shapes for the concept "flag" in NUS-Wide corpus.

Figure 6: Various visual patterns need a lot of training data

4.2 The Noisy, Ambiguous and Incomplete Tag Problems

We design the second experiment to evaluate the ability of our algorithm to tackle the noisy and ambiguous and incomplete tag problem in user-supplied tags. The baseline system is a keyword (tag) matching algorithm. That is, if the image contains the keyword in the tag list, the algorithm will regard it as relevant to the concept; otherwise, it is irrelevant. The results are shown in Figure 7.

We found that our approach achieves a relative improvement of 38% as compared to the keyword matching approach. This is because the Semantic Field approach selects and analyzes a group of tags as a whole, which provides essential context information and reduces the effects of noisy, ambiguous and incomplete tags.

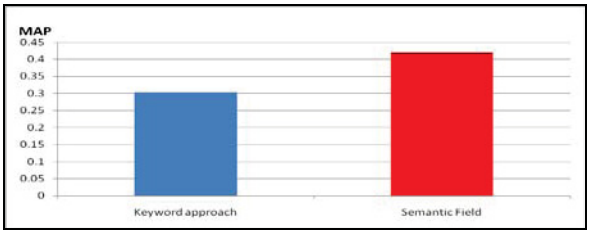


Figure 7: Comparison with keyword matching approach

For completeness, we also evaluate the system using the Equations (7) and (8) according to the top k images (k=1000, 2000, 3000, 4000, 5000).

$$P(tag) = \frac{\sum_{i=1}^{N} \frac{\#(p_i)}{\#(A_i)}}{N} \quad (7)$$

$$R(tag) = \frac{\sum_{i=1}^{N} \frac{\#(p_i)}{\#(T_i)}}{N} \quad (8)$$

We use p_i to represent the number of images with the target concept and A_i to represent the number of retrieved images for tag i. N denotes the number of different detected concepts (tags) in the ground truth set. In this corpus, the value of the N is 81. T_i is the number of the ground truth for a certain target concept.

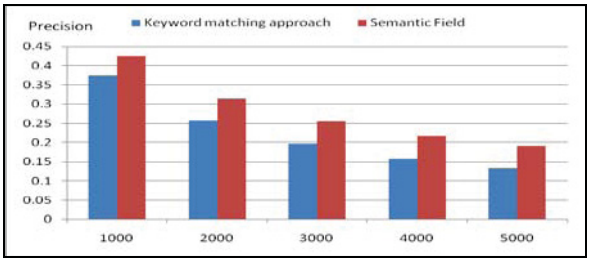


Figure 8: Comparison in precision on top-k image ranking. The x-axis indicates the value of k, while the y-axis shows the P(tag).

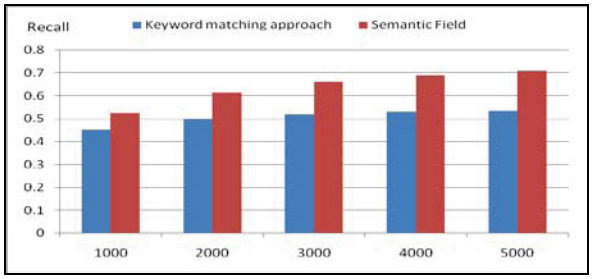


Figure 9: Comparison in recall on top-k image ranking. The x-axis indicates the value of k, while the y-axis shows the R(tag).

Figures 8 and 9 report the performance in precision and recall respectively. From the results, we find that our approach is better than that of the baseline system in both precision and recall. This is because on one hand the Semantic Field tackles the ambiguous and noisy tag problems so that we could improve the precision. On the other hand, the Semantic Field analysis includes many highly related tags, which tackle the incomplete tags problem so that it could improve the performance in recall.

4.3 Importance of Multi-source Information

Semantic Fields combine three information sources: WordNet, Wikipedia and the tag's co-occurrence information in the NUS-Wide corpus. We design the third experiment to evaluate the contribution of each information source. The results are shown in Figure 10.

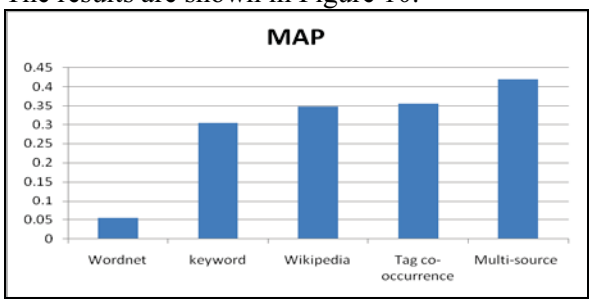


Figure 10: The comparison between using single information source and fusion of multiple information sources.

From Figure 10, we find that the performance of using WordNet alone obtains the worst result. This is because the number of tags carries the most common sense is limited and there are some noisy words in the description. For example, in Figure 2, the occurrence of the word "long" does not imply the occurrence of the concept "bear". Due to the lack of further information, using WordNet alone can hardly remove the noisy tag "long". The test result shows that such noisy information significantly degrade the performance of the system. This suggests the importance of incorperating other sources of informaiton to provide more complete information for the analysis.

We can also observe that using Wikipedia or tag co-occurrence shows comparatively better performance. This is because both information sources include abundance information for analysis. Thus, compared to the keyword-based approach, the performance of the systems shows around 17% improvement. Finally, fusing the three information sources results in the best MAP performance. This is because information from different sources complements each other and helps in reducing the effects of the noisy, ambiguous and incomplete tags.

5 Conclusion

In this paper, we proposed the use of Semantic Field to annotate web images. It could reduce the influences of noisy, ambiguous and incomplete tags so that the quality of the tags assigned to the web image can be improved. Our experiments showed that our approach is more robust and could achieve 38% improvement in MAP as compared to the learning-based and visual analysis approaches when there is sufficient text information. Also the fusion of multiple information sources could further boost the performance of the system.

The work is only the beginning. Future works include the followings. First, as multimedia data includes multiple modality features, how to fuse them to improve the performance of the system is an important problem. Second, current version of our algorithm only could identify one sense of the concept. How to distinguish among different senses of the concept is also an urgent task. Third, we will explore more semantic relations from Wordnet, Wikipedia and so on.

References

M. Ames and M. Naaman (2007), "Why We Tag: Motivations for Annotation in Mobile and online Media". In Proceedings of the SIGCHI confe-

rence on Human factors in computing systems, pp. 971 – 980.

C. F. Baker and C. J. Fillmore and J. B. Lowe (1998) "The Berkeley FrameNet Project", Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics pp. 86-90.

K. Bischoff, C. S. Firan, W. Nejdl, R. Paiu (2008), "Can All Tags be Used for Search", In Proceedings of the 17th ACM conference on Information and knowledge management, pp. 193-202.

T. S. Chua, J. H. Tang, R. C. Hong, H. J. Li, Z. P. Luo, and Y. T. Zheng (2009), "NUS-WIDE: A Real-World Web Image Database from National University of Singapore", ACM International Conference on Image and Video Retrieval.

B. M. David, A. Y. Ng and M. I. Jordan (2003), "Latent Dirichlet Allocation", Journal of Machine Learning Research 3: 993-1022.

P. Duygulu and K. Barnard (2002), "Object recognition as machine translation: learning a lexicon for a fixed image vocabulary", In Proceedings of the 7th European Conference on Computer Vision, 4: 97-112.

W. A. Gale and K. Church and D. Yarowsky (1992), "A method for disambiguating word sense in a corpus". Computers and the Humanities. 26 pp. 415-439.

S. H. Gao, L. T. Chia and X. G. Cheng, (2009) "Understanding Tag-Cloud and Visual Features for Better Annotation of Concepts in NUS-Wide DataBase" , In Proceedings of WSMC 2009.

M. F. Garrett (1992), "Lexical Retrieval Processes: Semantic Filed Effects", in Lehrer and Kittay Eds. Frames, Fields and Contrasts: New Essays in Semantic and Lexical Organization. pp. 377-396 Hillsdale: Lawrence Erlbaum.

R. E. Grandy (1992), "Semantic Fields, Prototypes, and the Lexicon", in Lehrer and Kittay Eds. Frames, Fields and Contrasts: New Essays in Semantic and Lexical Organization. pp. 103-122 Hillsdale: Lawrence Erlbaum.

T. Hofmann (1999), "Probabilitic Latent Semantic Indexing", In Proceedings of the 22rd Annual International SIGIR Conference on Research and Development in Information Retrieval.

J. Jeon, V. Lavrenko, and R. Manmatha (2003), "Automatic Image annotation and retrieval using cross-media relevance modes", In Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 119-126.

Y. Jin, L. Khan, L. Wang and M. Awad (2005), " Image Annnotations by Combining multiple Evidence & WordNet", In Proceedings of the ACM Multimedia Conference, pp. 706-715.

D. Jurafsky and J. H. Martin (2000), "Speech and language processing", published by Prentice-Hall Inc.

L. S. Kennedy, S. F. Chang and I. V. Kozintsev (2006), "To search or To Label", In Proceedings of MIR 2006, pp. 249-258.

R. M. V. Lavrenko and J. Jeon (2003), " A model for learning the semantic of pictures", In Proceedings of the 17th Annual Conference on Neural Information Processing Systems.

C. Manning and H. Schutze (1999). "Foundations of Statistical Natural Language Processing". MIT Press, Cambridge, MA.

K. Matusiak (2006), "Towards user-centered indexing in digitial image collections", OCLC systems and Services, 22(4): pp. 283-298.

R. Navigli (2009), "Word Sense Disambiguation: A Survey", ACM Computing Surveys, Vol. 41, No. 2. Article 10.

N. C. Rowe (1994) "Inferring depictions in natural language captions for efficient access to picture data", Information Process & Management Vol. 30 No 3. pp. 379-388.

G. Wang, T. S. Chua and Y. C. Wang (2003), "Extracting Key Semantic Terms from Chinese Speech Query for Web Searches". In proceeding of 41st Annual Meeting of the Association for Computational Linguistics pp. 248-255.

G. Wang, T. S. Chua, M. Zhao (2008), "Exploring Knowledge of Sub-domain in a Multi-resolution Bootstrapping Framework for Concept Detection in News Video", In Proceeding of the 16th ACM international Conference on Multimedia. pp. 249-258.

Merriam Webster Online dictionary (2010), Available at http://www.merriam-webster.com/