Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 3 (2011) 420–425

Procedia Computer Science

www.elsevier.com/locate/procedia

WCIT-2010

SemRank: ranking refinement strategy by using the semantic intensity

Nida Aslam [a] *, Irfan Ullah [ab], Jonathan Loo [a], RoohUllah [c], Martin Loomes [a]

[a]*School of Engineering & Information Sciences, Middlesex University, The Burroughs London, NW4 4BT,UK*
[b]*Department of Information Technology, Kohat University of Science & Technology, Kohat, 26000, Pakistan*
[c]*Departmentof Computer & Information Sciences,Universiti Teknologi Petronas, Bandar Seri Iskandar, Malysia*

Abstract

The ubiquity of the multimedia has raised a need for the system that can store, manage, structured the multimedia data in such a way that it can be retrieved intelligently. One of the current issues in media management or data mining research is ranking of retrieved documents. Ranking is one of the provocative problems for information retrieval systems. Given a user query comes up with the millions of relevant results but if the ranking function cannot rank it according to the relevancy than all results are just obsolete. However, the current ranking techniques are in the level of keyword matching. The ranking among the results is usually done by using the term frequency. This paper is concerned with ranking the document relying merely on the rich semantic inside the document instead of the contents. Our proposed ranking refinement strategy known as SemRank, rank the document based on the semantic intensity. Our approach has been applied on the open benchmark LabelMe dataset and compared against one of the well known ranking model i.e. Vector Space Model (VSM). The experimental results depicts that our approach has achieved significant improvement in retrieval performance over the state of the art ranking methods.

Selection and/or peer-review under responsibility of the Guest Editor.
*Keywords:*Ranking refinement; Semantic gap; Semantic Intensity; Multimedia Retreival.

1. Introduction

An increasing immensity of procurable digital data online as well as offline has simulated recent research into digital data mining, data management, data filtering and information retrieval. Due to omnipresence of these data, acquisition becomes a bottleneck. So there is an urge for the efficient and effective retrieval techniques. Merely finding the relevant information is not the only task of IR systems. Instead the IR systems are supposed to retrieve the relevant information as well as rank or organize according to its degree of relevancy with the given query.

Ranking is one of the intriguing issues in the IR systems. Ranking deals with sorting the retrieved results according to the relevancy with the given query. However, the result is the combination of the relevant as well as irrelevant data. The relevant document may have different degree of relevancy. The relevancy degree is defined as a "*function that determines the degree of semantic relatedness between the query and the retrieved results*". To

*NidaAslam, Tel: (+44) 07916860919
Email Address: aslam.neda@gmail.com

1877-0509
doi:10.1016/j.procs.2010.12.070

achieve high precision the relevant document must be top ranked. Retrieving the relevant information without appropriate ranking is obsolete.

The main stumbling block in ranking is to classify which documents are relevant and which are irrelevant.Existing ranking techniques mostly rely on keywords to judge the relevancy of the data with the given query. The relevancy was defined in terms of number of times the words that is in the query appear in the document i.e. term frequency. The document with the greater term frequency will be top ranked. The current techniques mostly rely on the keyword matching technique for finding the document relevancy with the query. But unfortunately the keywords alone cannot capture the entire semantics behind the query. The systems works well for simple object based queries. However, for the complex queries it's trivial and leads to the poor retrieval performance. This is the one of the main handicap of the traditional IR systems.

In order to achieve effective retrieval performance, instead of using the keyword or text matching technique for ranking, it must be done by exploring the intended meaning behind the group of words or keyword. There is a demand for the system that can rank the output by considering multiple features instead of single feature for exploring the semantics.

To tackle this problem, we propose a novel ranking strategy known as SemRank, which rank the retrieved results on the basis of the Semantic Intensity (SI), which is the *"concept dominancy factor, the greater SI value of the image will have greater relevancy with the query".* The inspiration for the SemRank is that retrieving the relevant information is not a difficult task for the state of the art IR systems but ranking of the required document is still an open challenge. We focus on improving the precision of the IR system by ranking the documents on the semantic similarity between the retrieved document and the user query. Our method, rank the result on the basis of the semantic dominance of the concept in the retrieved images. Based on the semantic intensity the retrieved documents are then ranked.

In the following section, we explore the existing state of the art ranking strategies. In section 3 we define our own proposed model. In section 4 we present our experimental setup in terms of the dataset as well as evaluation measure used. We compare our proposed result with one of the traditional model and present our result. Finally section 5 conclude the paper and discusses the results.

2. State of the art ranking strategies:

Extracting the relevant information from the corpus and then rank the information according to the relevancy order is one of the main functions in the IR systems. The ranking area in the data mining and IR has already been investigated by many researchers by assigning the calculating the frequency of the term in the query and the frequency of the query terms in the document, assigning the weights to the objects etc. It is worth saying that a true ranking strategy is the one in which the relevant documents comes before the irrelevant and less relevant ones.

Over the last few decades many IR models such as Boolean models, statistical and probabilistic models have been proposed [1]. In the Boolean model a document and a query is defined as the set of keywords. The relevance of the result or document with the query can be judge by using the Boolean operators. The Boolean model works well for the simple queries while it fails for the complex queries. The Boolean model assigns equal weights to all the relevant documents. This results in the difficulty of ranking the most relevant one than the less relevant. The statistical model represents the query and the document as a bag of words. The similarity between the query and the document is calculated on the occurrence frequencies. The probabilistic models also known as inference network calculate the relevancy of the document by using the probabilistic techniques. Rank the document by calculating the ratio between the relevant as well as the irrelevant one.

Much effort has been placed on the development of ranking startegies including RankBoost [2,3], RankNet[4], ListNet [5] , Page Rank [6], Vector Space Model (VSM) [7], iRANK (Interactive Ranking) [8], fRank [9] ,PPRank (Predict Popularity rank) [10] , Ada Boost (Adaptive Boosting) [11] , HostRank [12], topical PageRank [13], Quantum Probability Ranking Principle (QPRP) [18] etc. RankBoost uses the boosting approach for combining the preferences. RankNet uses the gradient descent algorithm to train the neural network model for ranking. ListNet uses the probabilistic approach for ranking. It uses a list wise approach and used objects as an instance. QPRP has been proposed to remove the document dependency problem of Probability Ranking Principle (PRP). The QPRP captures the dependency between the documents by means of quantum inference.

A learning algorithm on the basis of the support vector machine (SVM) has been developed for ranking known as Ranking SVM [14].Support vector machine (SVM) is a machine learning technique for ranking. Different researchers are trying to optimize the state of the art techniques like [15] proposes an algorithms to remove the hinge loss on SVM.A new learning strategy has been proposed known as learning to rank (LTR) it uses several document features [16]. LTR selects appropriate ranking function for each query. The inspiration of LTR is, it is not necessary that a ranking function which works well for the single query will work well for all the other set of queries. Different ranking function suits different queries. The ranking fusion technique has been also used to make the significant improvement in retrieval of hand writing recognition systems [17]. All these approaches aim at producing the efficient ranking algorithm in order to optimize the retrieval performance.

The Vector Space Model (VSM) is one of the well-known traditional retrieval models. That uses the bag of words approach for the text retrieval. While an image can be represented as a vector in N dimensional feature space. The query and image is represented as weighted terms which can be further used for ranking. Analogy to the text retrieval term frequency and inverse document frequency can be used to assign the weight to the query and image. The similarity between the image and the query is calculated by using the cosine measure.Limitation of the vector space model is that it focuses on the frequencies of the terms that are tagged with the image during annotation while doesn't consider the data inside the image. The vector space model relies on the text matching technique and is unable to consider the structural information. However, sometimes the cosine similarity between the query and the image is high but the semantic similarity between the image and query is low.

In the above many approaches to ranking are discussed which have been used for the text as well as the image retrieval. Although the area related to the text retrieval are matured but image retrieval is worth investigating. Most of these techniques retrieved and ranked the images on the basis of visual similarity. But still the precision of the system is low because the visual similarity is not the semantic similarity.

3. Proposed Semantic Ranking Framework:

Our line of research focuses on the ranking on the basis of the semantics not on the basis of frequency comparison between the query and the documents. We envision that in order to achieve the effective retrieval performance semantic similarity should be consider instead of the visual or the textual similarity between the query and the information obtain from the tags attach with the image i.e. annotation. In our research, we exploit the Semantic Intensity (SI) for ranking the image, we have implemented a semantic ranking strategy known as SemRank on the LabelMe dataset which is open source dataset available for academic and research, the object in the image in LabelMe dataset is represented by a set of points known as polygon, the area of the irregular polygon can be calculated by

$$A_{poly} = \frac{1}{2}\sum_{i=0}^{n-1}(x_i y_{i+1} - x_{i+1} y_i) \quad (1)$$

The Concept Dominancy for the given object can be calculated as

$$CD = \frac{A_{poly}}{I_s} \quad (2)$$

Where $I_s = h * w$, represents size of the image.The greater the SI value of the image relevant to the query, the higher is the rank.

Let $Q = \{t_1, t_2, \ldots, t_n\} = \sum_{i=1}^{n} t_i$ be the query with set of terms *'t'*, while Q' is the expanded or enhanced query with their semantic similarity values

$Q' = \{(t,SS)_1, (t,SS)_2, \ldots, (t,SS)_2\} = \sum_{i=1}^{n} (t,SS)_i$

The system must returns a subset of images C' from the corpus C, where C is set of images with their annotation represented by the following equation

$$C = \{I_1, I_2, \ldots, I_m\} = \bigcup_{j=1}^{m} I_j \quad (3)$$

Where $I = \{O_1, O_2, \ldots, O_z\} = \sum_{k=1}^{z} O_k$, then equation (3) become

$$C = \{I_1, I_2, \ldots, I_m\} = \bigcup_{j=1}^{m} (\sum_{k=1}^{z} O_k)_j \quad (4)$$

Procedure: SemRank

Input:

$Q' = \sum_{i=1}^{T'} (K', SS)_i$

$C = \sum_{j=1}^{n} (\sum_{x=1}^{m} O_x)_j$

Output:

$R' = SemRank\ Result$

Method:

1. Applying enhanced query on the corpus C
   $R \leftarrow Q' \cap C$, where $R \in Q' \cap C' and\ C' \leq C$
2. Foreach $C'.Image$ in $R.C'$
   a. Foreach $C'.Image.Object$ in $R.C'.Image$
      i. Calculate object dominancy OD for each object
      $$OD \leftarrow \frac{1}{2} \sum_{i=0}^{n-1} (x_i y_{i+1} - x_{i+1} y_i)$$
      ii. Calculate the concept dominancy of each concept tag with the object in the image
      $CD \leftarrow \frac{OD}{I_s}$, where $I_s = h * w\ of\ the\ image$
      iii. Calculate the Semantic Intensity (SI) for concepts relevant to the query
      $SI \leftarrow R.Q'.SS * SI$, where $R.Q'.SS$ is the semantic similarity value of each query term
   b. Calculate $netSI$ at $R.C'.Image$ level
   $R.netSI \leftarrow \sum_{i=1}^{n} (SI)_i$, where *n* is the number of concept tag with object per image
3. Sort the result in descending order

$R' \leftarrow sort(R.netSI, Descending)$

4. Experimental Study:

A comprehensive empirical performancestudy, using both Vector Space Model and SemRank has been made. The experiments were conducted on some of the categories from the LabelMe 31.8 GB dataset. Which contain total of 181,983 images, 56,943 annotated images and 125,040 images are still not annotated [19].The study is made with the objective to test the result of the proposed method against the traditional IR model i.e. VSM. Several experiments were conducted using different set of queries like keyword based queries which may either single concept or multi-concept and multi-word queries i.e. multi-word multi concept etc. From the last few decades the researchers are mostly using two of the well-known properties for measuring the retrieval performance i.e. Precision and the recall. Precision is the ratio of the retrieved documents that are relevant while recall is the ratio of the relevant documents retrieved by the system.

The result shows the P@10 i.e. precision of the top 10 retrieved images for the different types of queries. Figure 1 shows the precision of different type queries by using VSM and SemRank, figure 2 shows the overall precision recall graph. All these results depicts that SemRank shows effective improvement over the VSM.

Figure 1. Precision @10for VSM and SemRank

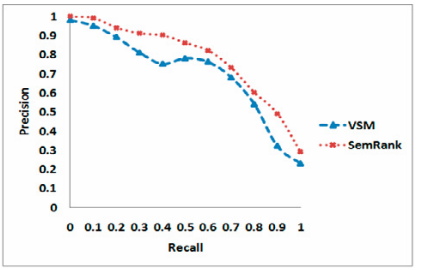

Figure 2. Precision-Recall comparison of SemRank performance over VSM

5. Conclusion and Future Work:

In conclusion we would like to accentuate that 100% retrieval performance is an exceedingly hard dilemma to achieve. The main problem lies in the deed that we don't totally formularize the term relevant and irrelevant. In this paper, we have proposed a new ranking strategy known as SemRank which uses the SI measure to calculate the image relevancy weights against the query. It has an advantage that it can employ the semantics inside the image and the query in determining the ranking order. We have compared our model with the Vector Space Model (VSM). Experimental results showed that SemRank approach has better retrieval performance than the VSM. We believe that considering the Semantic Intensities of the images enhance the precision of the IR systems. In future, we plan to exercise our approach on other image as well as on video datasets.

References

1. Baeza-Yates, R., and Ribeiro-Neto, B. Modern Information Retrieval. Addison Wesley, 1999.
2. Freund, Y., Iyer, R., Schapire, R. E., and Singer, Y. An efficient boosting algorithm for combining preferences. Journal of Machine Learning Research, 4: 933-969. 2003.
3. C. Rudin, R. Schapire, Margin-based ranking and an equivalence between AdaBoost and RankBoost, Journal of Machine Learning Research, 10: 2193—2232, 2009.
4. Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M.Hamilton, N., and Hullender, G. Learning to rank using gradient descent. In Proceedings of the 22nd International Conference on Machine Learning (Bonn, Germany. 2005).
5. Cao, Z., Qin, T., Liu, T. Y., Tsai, M. F., and Li, H. Learning to rank: From pairwise approach to listwise approach. In Proceedings of the 24th International Conference on Machine Learning (Corvallis, OR. 2007).
6. Yushi Jing, Shumeet Baluja (2008), "PageRank for Product Image Search", ACM.
7. G. Salton, A. Wong, and C. S. Yang (1975), "A Vector Space Model for Automatic Indexing," Communications of the ACM.
8. Furu Wei1,Wenjie Li, Wei Wang ,"iRANK: An Interactive Ranking Framework and Its Application in Query-Focused Summarization", 2009 , ACM.
9. Ming-Feng Tsai, Tie-Yan Liu, Tao Qin, Hsin-Hsi Chen, Wei-Ying Ma. FRank: A Ranking Method with Fidelity Loss, ACM , 2007.
10. Crammer, K., and Singer, Y. PRanking with ranking. Advances in Neural Information Processing Systems, 14: 641-647. 2002.

11. Jun Xu, Hang Li, AdaRank: A Boosting Algorithm for Information Retrieval, Proc. of SIGIR 2007, 391-398.
12. Xue, G. R., Yang, Q., Zeng, H. J., Yu, Y., and Chen, Z. Exploring the hierarchical structure for link analysis. In Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (Salvador, Brazil. 2005).
13. Nie, L., Davison, B. D., and Qi, X. Topical link analysis for web search. In Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (Seattle, WA. 2006).
14. Y. Cao, J. Xu, T.-Y. Liu, H. Li, Y. Hunag, and H.W. Hon, Adapting ranking SVM to document retrieval, SIGIR 2006.
15. Agarwal, S., Collins, M.: Maximum margin ranking algorithms for information retrieval. In: Gurrin, C., et al. (eds.) ECIR 2010. LNCS, vol. 5993, pp. 332–343. Springer, Heidelberg (2010).
16. Peng, J., Macdonald, C., Ounis, I.: Learning to select a ranking function. In: Gurrin, C., et al. (eds.) ECIR 2010. LNCS, vol. 5993, pp. 114–126. Springer, Heidelberg (2010).
17. Pena Saldarriaga, S., Morin, E., Viard-Gaudin, C.: Ranking fusion methods applied to on-line handwriting information retrieval. In: Gurrin, C., et al. (eds.) ECIR 2010. LNCS, vol. 5993, pp. 253–264. Springer, Heidelberg (2010).
18. Zuccon, G., Azzopardi, L.: Using the quantum probability ranking principle to rank interdependent documents. In: Gurrin, C., et al. (eds.) ECIR 2010. LNCS, vol. 5993, pp. 357–369. Springer, Heidelberg (2010).
19. Statistics on 8th July, 2010.