韓國 컴퓨터情報學會 論文誌
第13卷 第4號, 2008. 7.

2008-13-4-2-4

멀티-큐 통합을 기반으로 WWW 영상의 자동 주석

신 성 윤*, 문 형 윤**, 이 양 원*

A WWW Images Automatic Annotation Based On Multi-cues Integration

Shin Seong-Yoon*, Moon Hyung-Yoon**, Rhee Yang-Won*

요 약

인터넷의 빠른 발전으로 현재 HTML 웹 페이지에 내장된 영상들은 눈에 띄게 두드러졌다. 내용을 묘사하고 주의를 끄는 놀랄만한 함수 때문에 영상들은 웹 페이지에서 사실상 중요하게 되었다. 모든 영상들은 가공할 만한 데이터베이스로 구성되어있다. 게다가, 영상들의 의미론적인 의미도 주변의 텍스트나 링크에 의해 잘 표현된다. 하지만 이들 영상의 소수들이 주요 구에 정확히 할당되고 주요 구들을 현재의 영상에 수작업으로 할당하는 것은 매우 어렵다. 따라서 주요 구들을 추출하는 절차의 자동화는 매우 바람직하다. 본 논문에서는 먼저 저수준 특징, 페이지 태그, 전체적인 단어 빈도수와 지역적 단어 빈도수를 기반으로 한 WWW 영상 주석 방법을 소개한다. 그리고 멀티-큐 통합 영상 주석 방법을 전개해 나간다. 또한 실험을 통하여 멀티-큐 영상 주석 방법이 다른 방법보다 우수함을 보여준다.

Abstract

As the rapid development of the Internet, the embedded images in HTML web pages nowadays become predominant. For its amazing function in describing the content and attracting attention, images become substantially important in web pages. All these images consist a considerable database. What's more, the semantic meanings of images are well presented by the surrounding text and links. But only a small minority of these images have precise assigned keyphrases, and manually assigning keyphrases to existing images is very laborious. Therefore it is highly desirable to automate the keyphrases extraction process. In this paper, we first introduce WWW image annotation methods, based on low level features, page tags, overall word frequency and local word frequency. Then we put forward our method of multi-cues integration image annotation. Also, show multi-cue image annotation method is more superior than other method through an experiment.

▸ Keyword : 영상(image), 멀티-큐(multi-cues), 주석(annotation),

• 제1저자 : 신성윤
• 접수일 : 2008. 4. 30, 심사일 : 2008. 5. 29, 심사완료일 : 2008. 7. 25.
* 군산대학교 컴퓨터정보공학과 교수 ** LG CNS 차장

Ⅰ. Introduction

With the development of the Internet and the relevant technologies, and the availability of image capturing devices such as digital cameras, image scanners, the usage of images in HTML web pages is now predominant. These images can enrich the content of web pages and enable users to get intuitionist understanding of the content. This large collection of digital images becomes an important source from which users can get their target images with interest. How to get the most relevant results to the search query becomes an important issue.

In the earlier image retrieval systems, images are annotated manually by text descriptors. There are two disadvantages with this approach. The first is a considerable level of human labor is required for manual annotation. The second is the annotation inaccuracy due to the subjectivity of human perception. To overcome these two disadvantages in text-based retrieval systems, content-based image retrieval was introduced[1]. In CBIR, images are indexed by their visual features, such as color, texture, shape. Though many sophisticated algorithms[2][3][4] have been designed to describe color, shape and texture features, these algorithms cannot adequately model image semantics and have many limitations when dealing with broad content images databases.

As to web-based image retrieval systems, the abundant semantic meanings, such as captions, summary of images, surrounding text and hyperlink, give a particular description to the content of images.

This paper puts forward an automatic annotation mechanism, based on the surrounding semantic environment. Taking the localization, global term frequency and local term frequency into account to get semantic keyphrases for images.

Ⅱ. Approach

In the image retrieval field, the results are general evaluated at the semantic level. It means that the satisfying results are highly relevant to the users' query keywords(not the others) at the semantic level. Semantic is of great importance to the performance of image retrieval systems. Recently content-based image retrieval systems try to reduce the "semantic gap" between the visual features and the richness of human semantics, but in general, there is no direct relation between high-level concepts(keywords, text descriptors) and low-level features(color, texture, shape). Therefore, the result images of CBIR are only visually similar to the query images, not relevant by semantic.

2.1 Tag

HTML, an initial of Hyper Text Markup Language, is the predominant markup language for web pages. It provides a means to describe the structure of text-based information in a document by denoting certain text as links, headings, paragraphs, lists, and so on and to supplement that text with interactive forms, embedded images, and other objects. HTML is written in the form of tags, surrounded by angle brackets. HTML can also describe, to some degree, the appearance and semantics of a document, and can include embedded scripting language code(such as JavaScript) which can affect the behavior of web browsers and other HTML processors.

Cyber text is structured text[5] HTML documents use TAGs to organize the structure. The content of several TAGs has semantic relation to the image. Below in the Table 1, we list some text with different semantic meanings, ordered by their importance.

표 1. 태그 묘사
Table 1. Tag description

Tag	Description
Image Filename	Filename contains many important cues. The absolute URL, which has more information, is more important than the relative URL. Unfortunately, filenames usually are abbreviation or irrelevant description.
Caption	Usually, images have their captions which are around the images.
ALT	In HTML documents, images general have the ALT tag. It is used to describe content of images briefly while the images cannot show up. For example, "ThinkPad T43 A41" is used to describe the images about ThinkPad Computers.
HTML Title	HTML title frequently contains information about embedded images.
Hyperlink	Text of hyperlinks point out tips about hyperlinks.
Other text	Surrounding text maybe has something to do with images, but the relativity usually is weak.

As of the deletion of captions and the big noise of hyperlinks and other text, these text are determined by TF-IDF. The rest one will be weighted. Assume weights as follow, images filename - ω, ALT - τ, HTML title - λ, other - δ, $\omega < \tau < \lambda < \delta$.

Therefore, weight of some word $Wt = \omega \mid \tau \lambda \mid \delta$. Then we get the normalized weight as Eq. 1:

$$W_{tags}(t,\bar{d}) = \frac{W_t}{\sqrt{\sum_{t\in\bar{d}} W_t^2}} \quad \cdots\cdots\cdots\cdots \text{(Eq. 1)}$$

$\bar{d}$ is the current document.

2.2 TF-IDF

In practice, many images in HTML pages are not described as discussed above. The related text contents are organized mussily, but as to the whole document, we can still get the key information from the disordered structure. In the field of text classification, the vector space model is widely used to index text. Text features are tokens in text, so text can be indicated as feature vector $d=(t_1, t_2, \ldots, tn)$, t_i is the corresponding token weight. Feature selection actually is to select a proper subset $T'=\{t_1, t_2, \ldots, t_s'\}$ from the feature set $T=\{t_1, t_2, \ldots, ts\}$ $(s' \subset s)$. In text classification, these statistics are usually used for feature selection: Term Frequency, Document Frequency, Term Entropy, Multi2Information, Information Gain, Chi2square, Term Strength, Expected Cross Entropy, Weight of Evidence for Text, Odds Ratio. This paper mainly uses TF and DF[6] together for feature selection.

Term Frequency : The term frequency in the given document is simply the number of times a given term tk appears in that document. Intuitively, the more the feature appears, the more it contribute to the classification. Most original features are low frequency, so the proper threshold is effective to filter the low frequency features. The well-proportioned high frequency features in text is limited in classification. So, when used in text indexing, TF generally delete some low frequency features directly [7, 8].

Document Frequency : (The document frequency is the number of times a text which contains the token tk) The document frequency is a measure of the general importance of the term (obtained by dividing the number of all documents by the number of documents containing the term). It assumes that the low term frequency is not important in classification, or even is noise. If some low term frequency appears in some certain text, the feature will be filtered falsely. It has been researched that, TF together with IDF get the satisfying result in feature selection.

To put everything in a nut, TF is the number of times that a word appears in that document. If the word is very important to that article, it appears time and again, so it get a high TF weight. But these words which have high TF weight sometime

are not keyphrases of that document. For instance, some words appear frequently in many documents, so they cannot be distinguished keyphrases for certain documents. As of these, we use DF to indicate the feature, and get the IDF (Inverse Document Frequency) based on DF.

For the moment, there are many TF-IDF formulas, we take one universal formula described in Eq. .2 :

$$W_{tfidf}(t,\bar{d}) = \frac{tf(t,\bar{d})\log(N/n_t + 0.01)}{\sqrt{\sum_{t\in\bar{d}}[tf(t,\bar{d})\log(N/n_t + 0.01)]^2}} \quad \cdots\cdots\cdots \text{(Eq. 2)}$$

$W_{tfidf}(t,\bar{d})$ is the weight of t in text d. $tf(t,\bar{d})$ is the term frequency of t in text d. N is the total number of trained text. ni is the number of times that t appears in trained text. Denominator is the normalized factor.

The more bigger the weight of t, the more special the word is, in practice, we use the words with biggest weight as keyphrases of that document.

2.3 S-R

The associative nature of the Web has been under-exploited so far. In this paper, an extra feature was added that takes advantage of any semantic similarity that may exist between linked web documents.

The concept behind hypertext is that text content (or other media, in fact) is connected by associations or 'links' from document to document, forming a directed graph structure. The associations will usually (although not always) be based on some semantic similarity or relevance (of varying strength) between two documents.

The link structure of web documents is included by introducing the "Semantic Ratio"(SR) [8] feature. SR is similar to the TF-IDF feature, in that it is a frequency ratio. However, the SR of a phrase is calculated by dividing the number of occurrences of that phrase in the current document by the number of times it occurs in all documents directly linked to that document (i.e. those that are the targets of hyperlinks in the document). So we can get the formula described in Eq. 3.

$$SR(P,D) = \frac{Frequency-of-P-in-D}{Frequency-of-P-in-documents-linked-to-D} \quad \cdots \text{(Eq. 3)}$$

The reason behind including this feature is based on the intuition that the content of a web document is frequently semantically related to its neighbours (in the context of a graph structure, in other words, the documents linked to it) and that the subject matter (identified by the keyphrases) of the document is therefore in some way relative to their contents.

A low SR value (< 1) indicates that a potential keyphrase occurs more frequently in the document's neighbours than in the document itself. The higher the SR value, therefore, the more specific the phrase to this document, relative to its immediate surroundings. Note that this is different to the TF-IDF score as only a subset of the documents are used to compute it, namely those documents that from a localized sub graph with paths of length 1 from the original document.

Normally, a web page is semantically relevant to its linked web pages. Therefore, we can take linking text as reference to get keyphrases. In this paper, keyphrases are indicated by the value of SR.

SR is a linking text based parameter to indicate the importance of a word in original text. Usually, the content of web page is relevant to that one which is corresponding to the hyperlinks(specially hyperlinks that nearby images or in the same structure). SR algorithm (Eq .4) takes advantage of this connotative relevancy to extract keyphrases.

$$SR(t,\bar{d}) = \frac{frequency(t,\bar{d})}{frequency(t,link(\bar{d}))+1} \quad \cdots\cdots\cdots\cdots \text{(Eq. 4)}$$

$frequency(t,\bar{d})$ is the frequency of token t in

current document d.

$frequency(t, link(\bar{d}))$ is the frequency of token t in the sub layers(linked to current document d) of current document d.

The weight of token t in document d. (Eq. 5) is indicated on the base of Eq. 4:

$$W_{sr}(t,\bar{d}) = \frac{SR(t,\bar{d})}{\sqrt{\sum_{r\in\bar{d}} SR(t,\bar{d})^2}} \quad \cdots\cdots\cdots\cdots \text{(Eq. 5)}$$

Denominator is the normalized factor

Initial analysis of the SR distribution in keyphrases suggests that phrases with extrema SR values are more likely to be keyphrases. In other words, phrases that appear frequently in surrounding documents (low SR) have high relevance for the document in question. Also, phrases that occur very rarely in surrounding documents (high SR) will also have high relevance, suggesting that they indicate a topic that is specific to the current document.

III. Multi-cues Integration Annotation

WWW images locate in structural, networking documents, so the importance of a word can be indicated by its location, frequency. There are two patterns for multi-cues integration annotation.

(1) Linear Integration

In this paper, we use the below multi-linear integration formula Eq. 6 to compute word weight in documents:

$$W(t,\bar{d}) = \alpha\ W_{tag}(t,\bar{d}) + \beta\ W_{tfidf}(t,\bar{d}) + \gamma\ W_{sr}(t,\bar{d}) \quad \cdots\cdots \text{(Eq. 6)}$$

The importance of token t in document d consists of three cues' weight :

TAG weight $W_{tag}(t,\bar{d})$, TF-IDF weight $W_{tfidf}(t,\bar{d})$, SR weight $W_{sr}(t,\bar{d})$, α, β, γ are influence of each weight. After the weight is computed, some (for instance, 4) are chose to be the annotation keyphrases to the image.

(2) Tactic Integration

Tactic integration use each cue to select keyphrases orderly. We generally apply TAG to narrow the keyphrases range down, then we apply TF-IDF and SR to filter in candidate set. Fig 1 is the flow chart of multi-cues tactic integration.

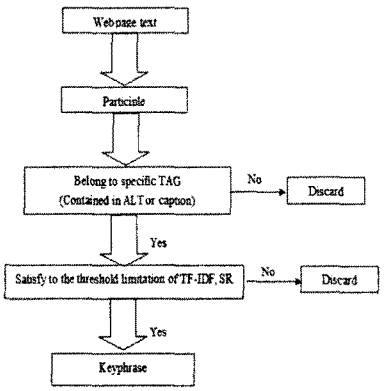

그림 1. 멀타-큐의 전술적 통합 플로 차트
Fig 1. Multi-cues tactic integration flow chart

TF-IDF threshold limitation can select threshold according to experiments, or simply order the tokens and take the first few ones as keyphrases.

IV. Experiments

The multi-cues integration based automatic image annotation algorithm was then tested on four web images corpora from online image search engines - Baidu、Google、MSN Live、Yahoo . These corpora were chosen because a sufficient number of documents in each site contained annotated keyphrases in the form of the Meta Keyword HTML tags and were therefore suitable for empirical tests on the accuracy of automatic keyphrase extraction.

First a crawler was used to fetch web-page content and hyperlinks. The WWW structure is a

directed graph with web-pages as notes and hyperlinks as borders. So the search progress of the crawler can be treated as a traverse of a directed graph. In this search progress, the crawler took the breadth-first strategy: searched all the hyperlinks in one web document and then the next layer until the last layer. This process is described below in Fig 2. The advantage of breadth-first strategy is that it can find the most shortest path between two web documents and lower down the visit frequency to the same server. Also its disadvantage is it cannot make a enough dig to deep web documents, or the search efficiency can be affected by the long time consuming dig progress.

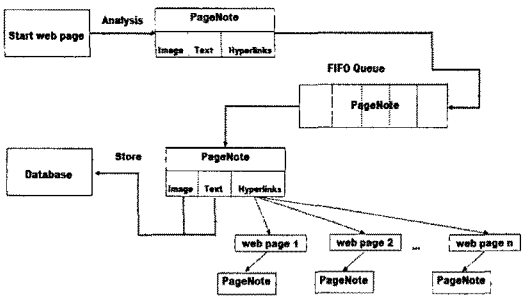

그림 2. Crawler 검색 절차 플로 차트
Fig 2. The Crawler Search Process Flow Chart

Then we got the necessary information attached to images from web-pages content. With these information and adapting the automatic keyphrase extraction algorithm discussed in Chapter 2, we can extract keyphrases to annotate these web images. This progress can be seen in Fig 3. Next step, we used an indexer to make an index with the help of these image annotations. The structure is described in Fig 3 below.

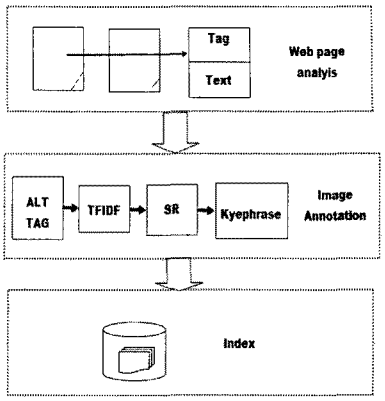

그림 3. 인덱스 모듈
Fig 3. Index Module

The indexer then extracted index entries from the abstract data of these analysed web pages. These index entries are actually keyphrases filtered by Meta tag, TF-IDF tag and SR tag. We selected the accessory Meta tags to the image and adapted participle program to get participles. Together with TF-IDF and SR, this multi-cues integration algorithm (Eq. .6) was then adapted to get keyphrases of the image. We usually take the first several ones with high W(t, d)(Eq. .6) as keyphrases, and the average number of correct keyphrases found in each corpus was recorded and is presented below in Fig.4.

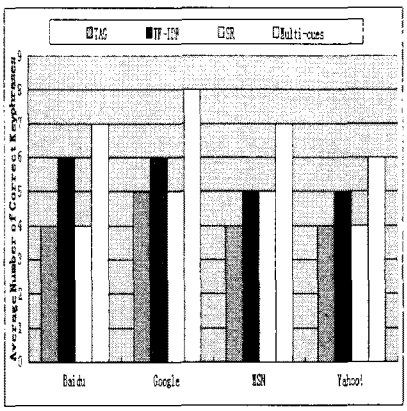

그림 4. TAG, TF-IDF, SR과 Multi-cue 기반 주석의 비교
Fig 4. Comparison of TAG, TF-IDF, SR and Multi-cues based Annotation

Fig. 4 shows that the integration of TAG, TF-IDF and SR feature improves the success of the multi-cues integration annotation algorithm by between 25% and 60% in these four test corpora.

V. Conclusion and Future Works

The multi-cues integration algorithm shows initial promise as an indicator of semantic keyphrases of the web images. The latent semantic automatic keyphrase extraction that causes the improvement with the usage of multi-cues is expected to be preferable. Future work therefore involves :

As the HTML5 or xHtml2 will come out in a few years, the organization of web information will be much more semantic. The source code of web pages will be more compact but with more highly semantic. So we can get the clear and ordered information we want from the tags much more easily and effectively. It will help to improve the accuracy and the efficiency.

Further analysis of the distribution of SR in keyphrases shows it is clear that the SR and TF-IDF features are not independent. Futhermore, while phrases with low TF-IDF are generally less likely to be keyphrases. This is not typically the case with SR. Therefore, the independence assumption will , in some cases result in a less accurate classification.

Experimentation with the SR feature is required in order to determine if a more suitable feature or number of features exist. In addition, the SR feature will be more generalized to take into account more distant documents than those directly related to the document in question, perhaps including a link-weighting or spreading-activation mechanism for retrieving the documents.

참고문헌

[1] Ritendra Datta, Dhiraj Joshi, Jia Li and James Z. Wang, "Image Retrieval: Ideas, Influences, and Trends of the New Age," ACM Computing Surveys, vol. 40, No. 2, pp. 1-77, 2008.

[2] J.Cox, et al. PicHunter: Bayesian relevance feedback for image retrieval [C]// Proceedings of 13th International Conference On Pattern Recognition, Vienna, pp. 361-369, 1996.

[3] Y. Rui, T. S. Huang, S. F. Chang, Image Retrieval: Current Technologies, Promising Directions and Open Issues [J]. Journal of Visual Communication and Image Representation, 10(4), pp. 39-62, 1999.

[4] Theo Gevers and Arnold Smeulders. Pictoseek: Combining color and shape invariant features for image retrieval. IEEE Transactions on image Processing, 9(1):102119, January 2000.

[5] Charles Frankel, Michael J. Swain, and Vassilis Athitsos, WebSeer: An Image Search Engine for the World Wide Web. Technical Report TR-96-14 [R], Chicago: University of Chicago, 1996.

[6] Salton, Gerard. Automatic text processing: the transformation, analysis, and retrieval of information by computer, Reading, Mass.Wokingham: Addison Wesley, 1998.

[7] Joachims T. A probabilistic analysis of the rocchio algorithm with TFIDF for text Categorization [C]// Proceedings of the 14th International Conference on Machine Learning, Nashvile, pp. 143-151, 1997.

[8] Apte C, Damerau F J, and Weiss SM, Automated learning of decision rules for text categorization [J]. ACM Transactions on Information Systems, pp. 233-251, 1994, 12.

[9] Daniel Kelleher, Saturnino Luz. Automatic Hypertext Keyphrase Detection [C]// International Joint Conferences on Artificial Intelligence (IJCAI), Edingurgh, pp. 1608-1609, 2005

저 자 소 개

신 성 윤

2003년 2월 군산대학교 컴퓨터과학과 이학박사

2006년~현재 군산대학교 컴퓨터정보과학과 교수

〈관심분야〉 비디오 인덱싱, 비디오 요약, 멀티미디어

문 형 윤

1997년 명지대학교 산업기술대학원 컴퓨터공학과 졸업(석사)

2007년 육군본부 전산장교

1997~현재 LG CNS 공공사업본부 차장

〈관심분야〉 영상처리, 컴퓨터비전

이 양 원

1994년 8월 숭실대학교 전자계산학과 공학박사

1986년~현재 군산대학교 컴퓨터정보과학과 교수

〈관심분야〉 모바일 프로그래밍, 텔레매틱스, 가상현실