Research on the Technology of Video Semantic Retrieval Based on Structured Semantic Strings

Jing Yuan, Quan Zheng, Zhijun Sun, and Song Wang

Department of Automation, University of Science and Technology of China,
Hefei, 230027, P.R. China
phoenix8@mail.ustc.edu.cn, qzheng@ustc.edu.cn

Abstract. To explore a concise and efficient approach to address video semantic retrieval problem, we creatively propose a concept of structured semantic string to realize video indexing and retrieval in this paper. Integrated with the technology of NLP (Natural Language Processing) and documents inverted index, our method could help represent information on sentence level so that computer could better understand videos' descriptions. In addition, expansion on synonyms and hypernyms contributes to comprehensive semantic expressing. Presently, the approach has been applied in our distributed video semantic retrieval system – Xunet, and it is proved to achieve high recall rate and precision rate.

Keywords: video semantic retrieval, NLP, semantic expanding.

1 Introduction

Facing a bunch of videos offered by video service websites, users depend on video search engine the same as they rely on general web search engine. How to provide videos in demand precisely is a hot problem puzzling both industry engineers and academic researchers. Apparently, the vital point to increase the retrieval effect is representing video information and search intensions semantically.

Nowadays, mainstream video retrieval technologies still rely on keywords matching. Text annotations attaching to videos help describe video information. These contents are then converted to words frequency vectors and stored in inverted indexes documents. Keywords refined from search contents are submitted to be retrieved in index [1]. Though the technology has been maturely utilized, it is not enough to accommodate users' demands any more. First, over-reliance on term frequency in a document to judge the similarity between a document and a term is not reasonable considering the interference of massive ineffective words repetitively appear in a document. Second, owing to lacking in analyzing relationships of words in a sentence, computer could only understand descriptions on words level rather than sentence level. That is to say, it may break up the comprehensive meaning of videos.

From the perspective of videos' low-level features analyzing, CBVQ (content-based visual query) technology which attracts many experts' attentions. Researchers are trying to depict videos through still key frame's characteristics such as color, texture and shape or motion trajectory in time domain like object tracking [2, 3]. There

Y. Wang and T. Li (Eds.): Foundations of Intelligent Systems, AISC 122, pp. 721–730.
springerlink.com

has already been some accomplished system on this theme. For instance, VideoQ of Columbia University [2], BilVideo of Bilkent University [3] and Multimedia Search and Retrieval System of IBM [4]. Nevertheless, it is difficult to use image's low-level features to reflect high-level semantic information. Besides, dimension disaster prohibits this idea smoothly applied in practical systems. As to object motion trajectory method, object shape recognition hasn't been perfectly solved yet. Therefore, many attempts on CBVQ solely focus on a specific field such as sports [5].

Ontology is a concept thriving in recent years and contributes to specific field to establish its structured and standard description. Admittedly, as a prospective method to describe a small world, ontology acquires W3C's adequate attention so that its standards such as RDF (Resource Description Framewrok) and OWL (Web Ontology Language) are formulated in a few years. These standards provide general terms for describing knowledge in domains [6]. Some researchers are trying to apply the concept of ontology in video information indexing and retrieval [7]. Unfortunately, domain dependency restricts ontology researching range to a tiny circle. Unlike a specified field such as pharmacy and marketing chain that possess their own criteria to refer to, it is improper to convert various video descriptions to a limited standard form deficient of vivid semantic information.

Just as drawbacks mentioned above, common users are fastidious with searching effects provided by common popular technology while novel methods are under research and many obstacles are intercepting a giant leap. In this paper, we maximize the usability of matured technology and creatively propose the concept of structured string to semantically present video information. We utilize the technology of NLP (Natural Language Processing) to process video description and generate them to structured semantic strings automatically. Inspired by the concept of ontology and a semantic image annotation tool Caliph, we divide processing results to several levels' strings. These steps could filter much of useless words or information and express sentences as 0, 1, 2 level of strings respectively. Afterward, synonyms and hypernym expansions of searching contents further construct comprehensive understand of searching intentions. If only descriptions comprise proper sentences, their meanings could be stored in inverted documents concisely and semantically.

The rest of this paper is organized as follows, section 2 introduces some related technologies and works, section 3 describes the semantic expression procedure, section 4 exhibits the experimental results of our approach, section 5 makes a short conclusion to the work we have done.

2 Related Works

It is undeniable that our work has borrowed some ideas from NLP, inverted indexes and Caliph's scene description, yet to our best knowledge, it is the first time to aggregate these elements in a Chinese video retrieval system to overcome semantic gap problem. These significant related technologies and works will be introduced in this section.

2.1 Caliph and Emir

Caliph and Emir are MPEG-7 based Java prototypes for digital photo and image annotation and retrieval. They support graph like annotation for semantic metadata and content based image retrieval using MPEG-7 descriptors [8]. Caliph offers a convenient user interface for users to annotate images manually. It describes a scene by a semantic graph which is consisted of entity nodes and relationships between these nodes. Emir could then retrieve these annotated MPEG-7 files in database or Lucene index. However, Caliph stipulates so many semantic relations that users' personal preference will cause different annotation results about a same image and it will definitely affect annotating precision and retrieval effect.

In our video retrieval system Xunet, we borrow the idea from Caliph to create our structured semantic strings. We also use the concept of 0, 1, 2 levels of strings to depict a scene, 0 level string represents a single entity, 1 level string contains two nodes and their relationship, 2 level string involves three nodes that represent subject, property and object respectively. These strings that are generated by NLP could explain a specific scene appropriately and could be treated as concrete terms when build index. What's more, that automatic process makes it possible for users to use natural language to describe a scene or an event averse the troublesome brought from manual annotation. Table 1 below shows an example of 3 different levels of structured semantic strings converted from a sentence. All sentences are translated from Chinese.

Table 1. Example of structured semantic strings

Form	Components
Sentence	I loveBeijing Tian'anmen Square.
0-Level Strings	{ I, love, Beijing-Tian'anmen-Square }
1-Level Strings	{AgentOf_I_love, ObjectOf_love_Beijing-Tian'anmen-Square}
2-Level Strings	{SVO_I_love_Beijing-Tian'anmen-Square}

2.2 Chinese Natural Language Processing

Unlike English which uses backspace as splitter between words, Chinese sentence possess no tag to distinguish different word. Both Chinese character and characters combination could represent a word. This feature results in the reliance on analyzer to split sentences automatically. In this paper, we make use of Chinese Analyzer which is created by Chinese Academy of Sciences to complete words splitting. Chinese Analyzer is a Java encapsulated tool package, which refers to related lexical library.

LTP (Language Technology Platform) is a language processing system framework designed by Harbin Institute of Technology. It uses XML-based texts to exhibit processing results and offers an effective bottom-up Chinese processing module, including some core technology as lexical analyzing and grammar recognition [9]. With the help of LTP's DLL application program interface and dependency tree, we add more functions such as sentences filtering and qualifiers extracting to form our own language processing procedure to cater for the demand of video description.

2.3 Lucene: Multiple Fields Retrieval

Lucene is an inverted full-text search engine tool which was originally developed by Doug Cutting and currently it is a subproject of the Apache Software Foundation project team. Lucene treats a string as a term and defines document consisted of many strings, thus a document could be represented as a vector of terms. Elements of these vectors are frequencies of these terms. Through vectors calculation, similarity between term and document could be computed by a specific formula defined in Lucene.

Moreover, Lucene provides practical interfaces to realize multiple fields' retrieval so that users could retrieve documents by their interesting theme. For example, we could separate description content to several themes and build up indices respectively. In this paper, we store keywords and structured semantic strings in different fields' index to compare their retrieval effects. Currently, we have realized three methods of retrieval: keywords retrieval, semantic graph retrieval and semantic structured string retrieval. To contrast their performances, we establish searching fields for each of them. We also provide three user interfaces to test retrieval effects.

2.4 Xunet Video Semantic Retrieval System

Our project named Xunet is a distributed video semantic retrieval system mainly consisted of video storage subsystem, video preprocessing subsystem, video retrieval subsystem and semantic processing subsystem. The architecture of this system is shown in Fig. 1. Reference [10] introduces the detailed system design and implementation, including deploying procedures. The subsystem in red rectangle in Fig. 1 is the major point we focus on in this paper.

As can be seen from Fig. 1, semantic processing subsystem, distributed indexing and retrieval subsystem are vital parts of this project which could directly determine the users' experience and evaluation. In the next section, we will emphasize on the semantic processing subsystem module and demonstrate its devotion to video semantic retrieval effect.

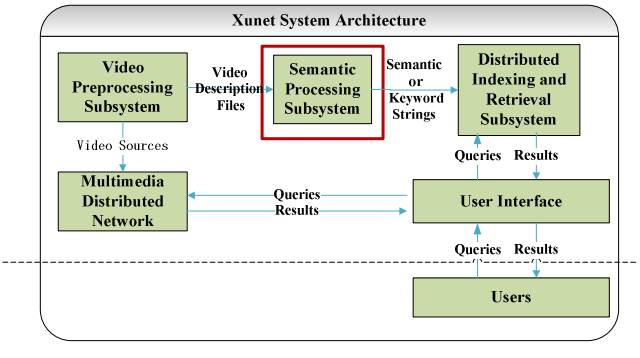

Fig. 1. Xunet system architecture

3 Video Semantic Retrieval

In this section, we will focus on detailed procedures of semantic conversion of video description. How to extract useful semantic segments which reflect video description's characteristic precisely? How to distinguish weights of different levels of semantic strings? What is a proper similarity calculating method? All these questions will be answered as following.

3.1 Semantic Structured Strings Formation

1) Semantic String Generating
To acquire structured semantic strings, we use semantic processing module to deal with it automatically. The process of this module is exhibited in Fig. 2 and described as follows.

a) Grammar Process Sub-module: Make use of LTP, mainly complete tasks of words splitting, part of speech recognition and determining grammatical relations between mutually dependent words (SBV, ATT, VOB etc.). *b) Sentence filtering sub-module:* Help judge and discard some useless or unconvertible sentences. *c) Main sentence extracting sub-module:* Remove some adverbials and attributes based on dependent relations left behind. *d) Sentence pattern recognition sub-module:* Analyze components of main sentence and determine the pattern of sentence. Mainly emphasize on five sentence patterns: declarative sentence, linked sentence, inversions, predicative sentence and complex sentence. Chinese sentence patterns are far more than these, while these five kinds are sufficient to basically represent video information. *e) Qualifiers extracting sub-module:* Conserve some potential useful qualifiers such as attributes and complements. *f) Semantic strings generating sub-module:* Express previous parsing results as structured semantic strings.

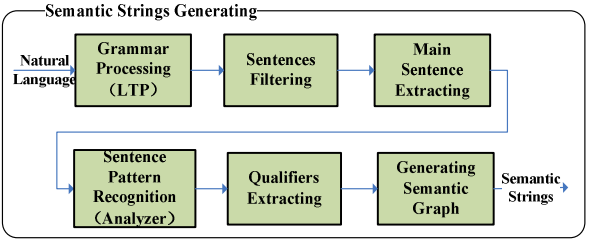

Fig. 2. Structured semantic strings generating

2) Relationship Selection
Among all three levels of generated semantic strings, it is evidently that 1-level strings and 2-level strings could better represent a sentence's integrity meaning. Since relations that connect entity nodes decide strings' form and structure, we have carefully chosen several relations which could not only express general videos' information but also simplify processing steps. Generally speaking, video content is consisted of scenes and events. Inferred from this opinion, we could draw the idea that entity

nodes could make up scenes and entity nodes companied with relations could represent specific events. If so, timing, location and agent or object could be treated as main elements of a video. As a matter of fact, according to the statistics of video descriptions we have acquired from video service websites by crawlers, these four elements indeed outweigh other elements. Therefore, we utilize these four relations (TimeOf_, LocationOf_, AgentOf_, ObjectOf_) to connect entity nodes. Experimental results in section 4 could demonstrate its reasonability.

3.2 Semantic Expansion

1) Single Node Expansion Procedure

Merely representing a sentence as a structured semantic string regardless of the existence of synonyms and relevant hypernyms may omit some familiar results when retrieve information among video descriptions. Only when all these relevant contents are taken into consideration, may the research intention be clearly represented. Hence we expand the searching content scope to synonyms and hypernyms. Our synonyms library derives from Chinese Synonyms Forests while hypernyms mapping table comes from Chinese WordNet, which was originally innovated by University of Princeton in English version and was than translated by Southeast University to Chinese version. Fig. 3 illustrates these two expansion procedure.

Each word has a serial number in Chinese Synonym Forest and a location number in WordNet. Though these two knowledge libraries own different encoding method, the function of their specific serial numbers or locations is to make similarity calculation more directly. There is always a minimum threshold of similarity or minimum level for expansion. Previous to the two expanded sets coming out and merging, words in the set should have already been filtered to meet threshold requirement so that irrelevant words won't infect final results.

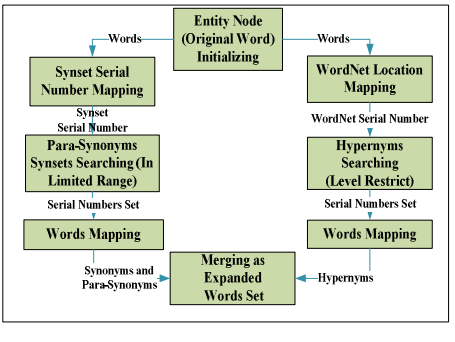

Fig. 3. Words expansion data flow diagram

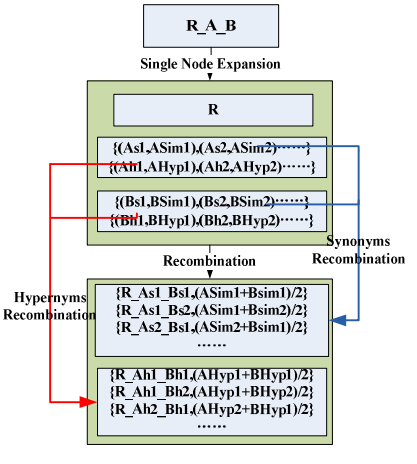

Fig. 4. Semantic strings' recombination

2) Structured Semantic Strings' Recombination

As the procedure introduced above, it only shows the expansion of a single entity nodes and the condition is only suitable for 0-level strings. Whereas for 1-level and 2-level semantic strings, after semantic expansion, more work should be done to recombine these expanded words as an entirety.

Take a 1-level semantic string's expansion recombination as example. Fig. 4 illustrates the detailed formation. Here we define R_A_B as a specific structured semantic string generated by natural language processing module mentioned above. In this figure, R means the relationship of entity node A and entity node B. Naming principle of A's relevant words is that As1 indicates one of A's synonyms, Ah1 indicates one of A's hypernyms, Asim1 means the similarity between As1 and A, AHyp1 means the hypernym level degree between Ah1 and A, and naming principle of B's relevant words could be inferred then. Red arrows sign A and B's hypernyms recombination, blue arrows sign A and B's synonyms recombination.

When word A and word B go through the expansion process, they could be transformed to two sets, one is synonyms set, the other is hypernyms set. These sets are comprised of pairs consisted of relevant strings and similarity or level degrees. When recombination completed, the fresh semantic string will own its specific weight value. Synonyms combination's weight equals to half of the sum of ASim and BSim, hypernyms combination's weight equals to half of the sum of AHyp and BHyp. Consequently, the expanded set is made of pairs of structured semantic string and its proprietary weight.

3.3 Weights Setting and Similarity Calculation

1) Level Weights Setting

Due to our retrieval mechanism based on inverted index, strings frequency in a document still determines retrieval results' matching degree. So it is unavoidable that 0-level semantic strings possess the largest probability to be retrieved. However, only 1-level strings and 2-level strings could evidently better exhibit a sentence's whole meaning. Large amount of 0-level strings involved in description will definitely spoil the retrieval accuracy. Take an example of a sentence which means "Clouds are floating." in English. It is possible that documents with high frequency of 0-level string "Clouds" or "floating" would acquire higher similarities and ranks while those documents containing "Cloud is floating" with one appearance are easily be omitted or ranked lower. With regard to this problem, we stipulate specific weight for each level's strings to enhance the importance of high level strings. Current level weights setting condition is listed in Table 2.

Table 2. Level weights setting

String's Level	Weights Parameter	Value
0-Level Strings	ZERO_CLASS	0.2
1-Level Strings	FIRST_CLASS	1.0
2-Level Strings	SECOND_CLASS	5.0

2) Inner Words Weights Setting

In the procedure of grammar analyzing, to maximize the information retaining, we sometimes bound adverbs or state qualifiers with nouns as an entity. For instance, in a structured semantic string which means "SVO_I_love_Beijing-Tian'anmen-Square" in English, "Beijing-Tian'anmen-Square" is a location entity and three words in this semantic string all contain important information. It is necessary to disassemble the original entity to an entity set. In this instance, original string will be transformed to a set consisted of three strings: "SVO_I_love_Beijing", "SVO_I_love_Tian'anmen" and "SVO_I_love_square". Thus location information could be separated and dispersed into several strings with inner words. To distinguish the original string and string with inner words, inner words weight should be considered. Table 3 shows the inner words weights setting. Such semantic strings with inner words should be multiplied by inner words weight after they multiplied by level weight.

Table 3. Inner words weights setting

String's Level	Weights Parameter	Value
Original String	ME_SIM	1
String With Inner Words	ME_INNER_SIM	0.2

3) Synonyms' Similarity Calculation

In order to discriminate synonyms and original words' matching degree, we use similarity calculation to determine the weight of a semantic string after synonyms expanding. Fig. 5 expresses synonyms forest structure and Chinese words' encoding formation.

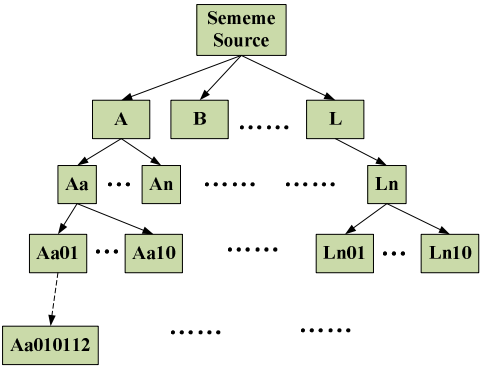

Fig. 5. Synonyms forest structure

Distance between words is a key standard to measure the relevance of two words. In many cases, directly computing the similarity is difficult so that distance between two words is always calculated at first step, and then it could be transformed to words' similarity [11]. Words distance is a real number in the range of [0-∞] and the distance between a word and its own is 0. In addition, there exists a principle, the greater the distance is, the lower the similarity will be.

As for two words $w1$ and $w2$, we express the distance between them as $Dis(w1,w2)$, and the similarity as $Sim(w1,w2)$. Then we define a simple transformation meeting above conditions:

$$Sim(w1, w2) = \frac{\alpha}{Dis(w1, w2) + \alpha} \tag{1}$$

In (1), α is an adjustable parameter, in this system, we set α as 1. Considering the forest structure of this synonyms library, we calculate the distance by shortest path between two word sememe's nodes. The similarity could be then generated by the formula above.

4 Experimental Results and Analysis

In this paper, we use web crawler to attain massive video descriptions as test data. These data come from 4 main Chinese video service websites and video themes involve in sports, news, movies and variety shows. By counting recall rate and precision rate, we could clearly illustrate the comparison of three level's structured strings' recall rate and precision rate in Table 4 and Table 5. Results of None Expanded Mode and Expanded Mode are also exhibited below.

Table 4. Recall rate

Strings' Levels	None Expanded	Both Expanded
0-Level	0.6094	0.6322
1-Level	0.6911	0.7311
3-Level	0.7455	0.7844

Table 5. Precision rate

Strings' Levels	None Expanded	Both Expanded
0-Level	0.6250	0.6417
1-Level	0.6350	0.6583
3-Level	0.6756	0.8316

From the statistics above, we can see that the higher the string's level is, the higher the recall rate and precision will be. In addition, synonyms and hypernyms expanded mode could provide a better performance than none expanded mode do. It is natural that users' experience will be enhanced by structured semantic strings' expression and its expansion results.

5 Conclusion and Future Works

We propose a new method to semantically retrieve video description in this paper. On one hand, after NLP parsing each sentence in documents, the structured strings containing grammar information will be indexed in the inverted indexing files. On the other hand, searching contents would experience NLP parsing and then be disassembled and expanded, when searching contents aggregate again as a searching set, it could be submitted and help complete a better searching effect. The approach has been applied in our video semantic retrieval system Xunet and performs well. Also,

the experiment tested on Xunet demonstrates the advance of our method by concrete statistics.

Admittedly, there exist some limitations of our approach. For example, sentences which can't be processed by NLP will be discarded and it may cause losing information, in addition, too large quantity of expansion will influence the searching effect. We will concentrate on these problems to balance the semantic expressing and searching efficiency in the future.

Acknowledgments. Our work was supported in part by the National Key Technology R&D Program (2008BAH28B04), "A New Generation of Video and Television Service System with Support for Cross-regional and Multi-operator". What's more, project "Content Distributed Service Commercialization Based on IPV6" (CNGI-09-03-14) hold by National Development and Reform Committee sponsored our work in the later period.

References

1. Lee, D.L., Huei, C., Seamons, K.: Document Ranking and the Vector-space Model. IEEE Transaction on Software 14, 67–75 (1997)
2. Chang, S.-F., Chen, W., Meng, H.J., Sundaram, H., And, D.: A Fully Automated Content-Based Video Search Engine Supporting Spatiotemporal Queries. In: IEEE Transactions on Circuits and Systems for Video Technology, pp. 602–615 (1998)
3. Baştan, M., Çam, H., Güdükbay, U., Ulusoy, O.: BilVideo-7: An MPEG-7 Compatible Video Indexing and Retrieval System. IEEE Transaction on Multimedia 17, 62–73 (2010)
4. Campbell, M., Haubold, A.: IBM research TRECVID-2006 video retrieval system. In: TREC Video Retrieval Evaluation Proceedings (2006)
5. Duan, L.-Y., Xu, M., Tian, Q., Xu, C.-S.: A Unified Framework for Semantic Shot Classification in Sports Video. IEEE Transactions on Multimedia 7(6), 1066–1083 (2005)
6. Shady, S., Fakhri, K., Mohamed, K.: Enhancing Search Engine Quality Using Concept-based Text Retrieval. In: International Conference on Web Intelligence, IEEE/WIC/ACM, pp. 26–32 (2007)
7. Ballan, L., Bertini, M., Del Bimbo, A., Serra, G.: Video Annotation and Retrieval Using Ontologies and Rule Learning. IEEE Transaction on Multimedia 17, 80–88 (2010)
8. Lux, M.: Caliph & Emir: MPEG-7 photo annotation and retrieval. In: Proceedings of the Seventeen ACM International Conference on Multimedia, Beijing, China, pp. 925–926 (2009), http://www.semanticmetadata.net/features/
9. Harbin Institute of Technology LTP website, http://ir.hit.edu.cn/
10. Zheng, Q., Zhou, Z.: An MPEG-7 Compatible Video Retrieval System with Support for Semantic Queries. In: International Conference on Consumer Electronics, Communications and Networks (CECNet), pp. 1035–1041 (2011)
11. Xu, S., Zhu, L.-J., Qiao, X.-D., Xue, C.-X.: A Novel Approach for Measuring Chinese Terms Semantic Similarity based on Pairwise Sequence Alignment. In: Fifth International Conference on Semantics, Knowledge and Grid, pp. 92–98 (2009)