International Journal of Computer (IJC)

ISSN 2307-4531

http://gssrr.org/index.php?journal=InternationalJournalOfComputer&page=index

SEMANTIC CLUSTERING WITH CONTEXT ONTOLOGY FOR INFORMATION RETRIEVAL SYSTEM

Thinn Lai Soe

[a] *University of Technology (Yatanarpon Cyber City)*

[a]*thinnlaisoe@gmail.com*

Abstract

Nowadays, there are so many increasing amount of information within world-wide web. For these increasing amounts of information, we need efficient and effective index structure when we have to find needed information. Most indexing techniques directly matched terms from the document and terms from query. But there is a problem when matching. That is most system doesn't consider the meaning of the words. A word can have more than one meaning. But most systems didn't consider the context (multiple meaning of a word). This paper presents how to construct an index structure using SSTC and context ontology that provides multiple meanings of a word. Context provides extra information to improve search result relevance. This paper produces context semantic cluster to provide indexing of search engine.

Keywords: indexing, context ontology, semantic suffix tree clustering (SSTC);

1. Introduction

Information Retrieval (IR for short) which finds information that actually need among large collection of documents is concerned with representing, searching, and manipulating large collections of electronic text and other human-language data [1]. Web search engines — Google, Bing, and others— are by far the most popular and heavily used IR services, providing access to up-to-date technical information, locating people and organizations, summarizing news and events, and simplifying comparison shopping[1]. The main aim of retrieval system is to provide most relevant documents to users. Therefore, giving relevant results, granting efficient and fast access is the main focus on the performance of retrieval system. The basic method of traditional IR is to find documents that contain the terms in the user query. Given a user query, one option is to scan the document database sequentially to find the documents that contain the query terms [1]. However, this method is obviously impractical for a large collection, such as the web. Another option is to build some data structures (called indices) from the document collection to speed up the search. IR model governs how a document and a query are represented and how the relevance of a document to a user query is defined [14]. There are four main models: Boolean model, vector space model, language model and probabilistic model [1][14]. But these models based on keyword or term matching, i.e., directly matches terms in the user query with those in the documents. If a user query uses different words from the words used in a document, the document will not be retrieved although it may be relevant because the document uses some synonyms of the words in the user query [1]. This causes low recall. For example, the words "doctor", "physicians", "surgeon" are synonyms in the context of "doctor". If the system doesn't consider these contexts of a word "doctor", we all don't have a chance to see many papers that use synonyms of the word that is used in users' query. This leads to context semantic indexing. Since there are typically many ways to specify a given concept (synonymy), the literal terms in a user's query might not match those of a relevant document. Moreover, most words have multiple meanings (polysemy), therefore terms in a user's query can literally match terms in irrelevant

documents [14]. An improved approach would permit users to retrieve information on the basis of a conceptual meaning.

Ontologies have been realized as the key technology to shaping and exploiting information for the effective management of knowledge and for the evolution of Semantic Web and its applications [15]. Ontology defines the terms and concepts (meaning) used to describe and represent an area of knowledge [5]. Context ontology defines a common vocabulary to share context information in a pervasive computing domain. For example, Figure 1 presents a part of simple ontology for celebrity David Beckham who is a footballer as well as brand name of accessories such as T-shirt, perfume, body lotion and shower gel etc.

This paper is organized as follows. Related works are presented in Section 2. Section 3, describes proposed system and advantages of the system are submitted in Section4 and then conclusion is in the last section, Section 5.

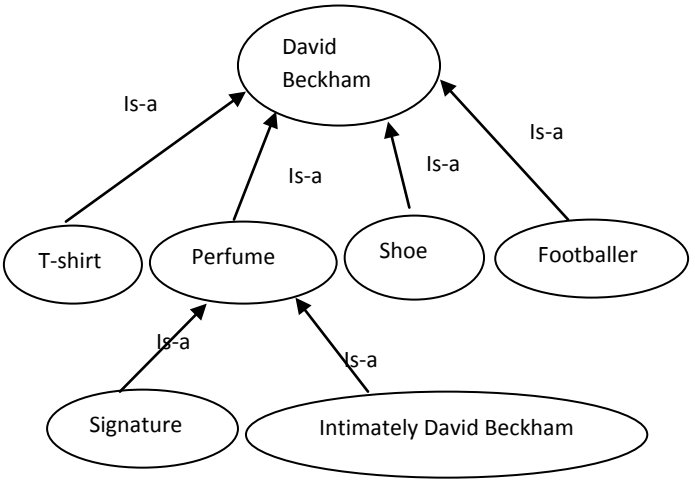

Figure1.Different context of David Beckham

Related Work

In this paper, a review of previous works on indexing is shown. There are many techniques that have been proposed already but such techniques produce inefficient and inaccurate results. In paper [2], the author proposed context based indexing using ontology. In this system, documents from the repository are first preprocessed and then, extract keyword with maximum frequency with the title. After that, extract context of the document using thesaurus and context ontology. A word which is not extracted cannot have a chance to become a context word. At that time, context of that word is not extracted from the document. The problem of this paper is that how to define the context of the word that is not chosen because of maximum frequency value.

Latent semantic indexing [3] is proposed by Deerwester et al, aims to add semantic recognition in retrieval. It uses a statistical technique, called singular value decomposition (SVD), to estimate this latent structure, and to remove the noise. The results of this decomposition are descriptions terms and documents based on the latent semantic structure derived from SVD [14]. This structure is also called the hidden concept space, which associates syntactically different but semantically similar terms and documents. But main drawback is the time complexity of SVD which is difficult to use for a large document collection such as the Web. Another drawback is that the concept space is not interpretable as its description consists of all numbers with little semantic meaning.

2. Proposed System

This paper submits the the model of how to construct a semantic index using the concept of ontology. Moreover, the system use clustering algorithm known as Semantic Suffix Tree Clustering (SSTC) and solving lack

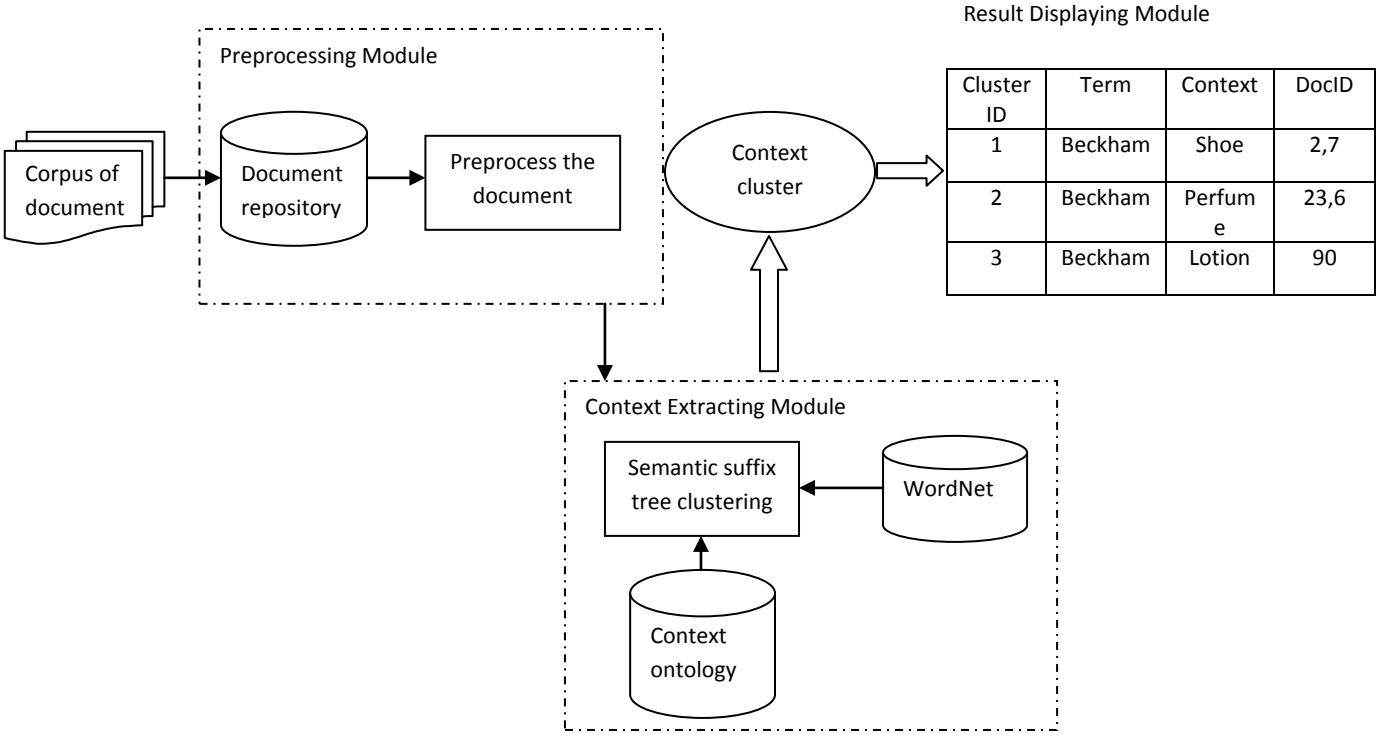

Figure2: Proposed System Architecture

of context which is a problem of SSTC. Context provides extra information to improve search result relevance. Context of a document can be extracted using the context ontology. The proposed architecture of the system is shown in above figure, Figure 2.

There are three modules in the system, the first is the preprocessing module, the second one is context extracting module and the last is the result displaying module. The modules of this system are described as below.

3.1. Preprocessing Module

Documents from the web are stored in the document repository. In the first module which is known as preprocessing module, documents from the repository are preprocessed. Preprocessing technique includes tokenizing the word, removing stop words, stemming the word and other preprocessing tasks for text are performed. For example "computer", "computing", and "compute" are reduced to "compute" [1]. Other preprocessing tasks include removing digits, breaking hyphens and punctuation marks and converting to either upper case or lower case.

3.2. Context Extracting Module

The document which is done preprocessing is passed into the context extracting module. The role of this module is the creation of clusters which are according to the context of the document. So, this module is the main component of the system. The system uses semantic suffix tree clustering to create clusters. Since SSTC cannot create context cluster, the system add context ontology so that SSTC with context ontology can create context clusters which provide index structure. The detail of this module is as follows.

3.2.1. Semantic Suffix Tree Clustering

Semantic suffix tree (SST) is a data structure that extends the suffix tree which requires using the suffix tree on the meaning of word strings not characters [4]. The meaning of word is an important property in semantic analysis. Semantic similarity equation is derived from the WordNet database by calculating the semantic similarity of word pairs. Therefore, semantic suffix tree use semantic similarity and string matching to create suffix tree. The equation which calculates semantic similarity is as follows. And semantic similarity measures the similarity deriving the synonyms sets from the WordNet, a lexical database [4].

SemSim (w_a,w_b) = 1 if $|\text{synset}(w_a) \cap \text{synset}(w_b)| \geq 1$
SemSim (w_a,w_b) = 0 otherwise

The cluster that has produced from the SSTC is the semantic cluster but it cannot know the context of that cluster. For this reason the system use context ontology to produce context clusters. The algorithm of the semantic suffix tree [4] is as follows:

Algorithm1: Construct Semantic Suffix Tree
1: input <-- set of String
2: output <-- semantic suffix tree
3: for each word (txt)
4: if root is empty then {
5: create a new node
6: and update position
7:} else{
8: add a new node into cn
9: do until pn = root {
10: if SemSim = 0 && no match then{
11: add a new node into pn
12: and update suffix link
13: update position
14 :} else {
15: update suffix link and update position}
16 :}
17 :}
18:}

3.2.2. Context ontology

Ontology can be viewed as the backbone to support various types of information management including IR, storage and sharing on web. It is used to reason about the entities within the domain, and may be used to describe the domain [6]. Context ontology include context of the term, that term which create context, their relationships among context and the term and other terms which support to make the context. Table 1 shows the terms which make the context and context of the terms. When we see "Signature and David Beckham" together in the document, the context of "David Beckham" does not mean either a footballer or name of body lotion. It only means the type of “Perfume” named "Signature Eau De by David Beckham". Context ontology is constructed so as to distinguish the context of the term like “David Beckham” using the concept of ontology. As we all know, ontology is a formal specification of a conceptualization or declarative representation of knowledge relevant to a particular domain [5]. Therefore, we can easily find the relationship between the contexts of the term, the main term and other terms that support finding the concepts of the term.

Table 1: Terms with context and the other terms that cause context

Context	Terms	Other Terms
Footballer	David Beckham	Manchester United
Perfume	David Beckham	Signature
Perfume	David Beckham	Intimately Beckham
Body lotion	David Beckham	Curious 6.0z
T-shirt	David Beckham	LA Galaxy Home

3.3. Result Displaying Module

This module presents the context cluster to user which differentiates the context of the term. Users can easily see the term whether it is the name of the song or it is the name of body lotion because of semantic suffix tree clustering and context ontology. The following table shows the final result of the documents.

Table 2: Results of the system

Cluster ID	Term	Context	DocID
1	David Beckham	Shoe	2,7
2	David Beckham	Perfume	23,6
3	David Beckham	Lotion	90

4. Advantages of the system

This section describes the benefits that get from this system. There are three advantages of the system describe here.

Firstly, if a system does not distinguish the cluster according to context, it does not have a chance to know whether the term is a name of perfume or it is a name of footballer, T-shirt etc when the system meets with the term "David Beckham". Since the proposed system has generated a context cluster, the system gets T-shirt clusters of David Beckham, perfume clusters of David Beckham etc. From this benefit, the proposed system gives a precise indexing structure because the system indexes these several documents according to the context cluster. The second advantage of the system is as follows. SSTC only solve the synonyms (which is one of information retrieval problem). The proposed system describes here, SSTC with context ontology, can solve the both synonyms and polysemys (two IR problems).

The third advantage of this proposed system is described below. Being used SSTC, the benefit of SSTC is that it generates overlap clusters. If a document is erroneously placed on wrong clusters, SSTC gives a chance to place that document in another true cluster.

5. Conclusion

Index construction is an important part in information retrieval system. This system constructs an index structure which is called context semantic index structure to support the applications related with information

retrieval. This index structure is constructed using the clustering method called semantic suffix tree clustering and context ontology. This system can overcome the two information retrieval problems called synonyms (multiple words with same meaning) and polysemy (a word with different multiple meanings). The proposed system produces precise and effective context semantic index than other indexing method.

Acknowledgement

I would like to thank my supervisor and all of my friends gratefully.

References

[1] Information Retrieval and Web Search (chapter 6) from Web Data Mining, Exploring Hyperlinks, Contents and Usage Data.
[2] Gupta P., and Sharma A.K., "Context based Indexing in Search Engines using Ontology", International Journal of Computer Application, 2010.
[3] S. Deerwester, S.T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 1990.
[4] Janruang, J., Guha, S.: Semantic Suffix Tree Clustering. In: DEIT 2011, IEEE, Bali, Indonesia (2011).
[5] Haibo Jia, Julian Newman, Huaglory Tianfield "A new Formal Concept Analysis based learning approach to Ontology building"
[6] http://en.wikipedia.org/wiki/Ontology (information_science)
[7] Janruang, J., Guha, S., "Applying Semantic Suffix Tree Clustering"
[8] Oren Zamir and Oren Etizioni, Web Document Clustering: A feasibility demonstration. In the proceedings of SIGR, 1998.
[9] C.Manning, P. Raghavan, and H.Schutze, "An introduction to information retrieval, "Cambridge, England: Cambridge University Press, 2009.
[10]Maxim Marynov, Boris Novikov, "An Indexing Algorithm for Text Retrieval", University of St.-Petersburg, Russia.
[11] E.W. Brown, J.P. Callan, W.B. Croft, and J.E.B. Moss. Supporting full-text information retrieval with a persistent object store,. In Proc. Intnl.Conf, on EDBT., 1994.
[12] Sajendra Kuar, Ram Kumar Rana, Pawan Singh, "A Semantic Query Transformation Approach Based on Ontology for Search Engine", International Journal on Computer Science and Engineering (IJCSE), May 2012.
[13] R.Baeza-Yates and B.Ribeiro-Neto. Modern Information Retrieval. Addison Wesley, 1999.
[14] N Chen, Technical Report 2006-505 “A survey of Indexing and Retrieval of Multimodal Documents: Text and Images”.
[15] K. Kotis, G. A. Vouros, K. Stergiou, Department of Information and Communication System Engineering, “Towards Automatic Merging of Domain Ontology: The HCONE- mearge approach.