NOVATEUR PUBLICATIONS
INTERNATIONAL JOURNAL OF INNOVATIONS IN ENGINEERING RESEARCH AND TECHNOLOGY [IJIERT]
ISSN: 2394-3696
VOLUME 5, ISSUE 3, Mar.-2018

INTENT-BASED DIVERSIFICATION FOR FUZZY KEYWORD SEARCH OVER XML DATA

SIJIN P
Department of Computer Science and Engineering, Bangalore University /University Visvesvaraya College of Engineering, Bangalore, India
psijin@gmail.com

ABSTRACT
The keyword queries over various data forms have wide attention nowadays. The query intention can easily be obtained by comparing the keyword with some query suggestions. An annotation process can be recommended to generate the structured meta-data for a document. In the proposed system the conceptualization and measure of co-occurrence count of a typed term has considered on the basis of semantic relatedness and similarity between terms. This should be useful for the retrieval of information by search query with short and vague keywords. Apart from this, using a fuzzy uncertainty function the fuzzy semantic of a query can be easily obtained. This will reveal the uncertainties among the co-occurring terms. The closest matching terms can be easily constructed by using the keyword similarity semantics. The edit distance and gram based pattern matching methods are used to check the closeness of keyword similarity of terms. The concept based clustering is used to hold the data in a multidimensional space. The proposed dimension reduction method reduces the cardinality of the result set in the direction of Eigen vectors calculated for the selected features. The outliers of the concept vector can set to the required level on the basis of concept density. The attributes and related features are stored as metadata information in XML files for more precise representation.

INDEX TERMS - concept, knowledge base, coherence, entity linking.

INTRODUCTION

In any knowledge database, which may contain millions of words, meaningful and redundant sentences, numeric data, abbreviations, and noise etc. The methods like Named Entity Recognition (NER), Topic models, Entity linking, Conceptualization, Opinion mining, Sentiment analysis, Summarization and Event detection are used to understand the semantic meaning of terms in a text on various contexts. The text segmentation result set can be extending with fuzzy matches of the search terms. The semantic meaning cannot preserve if considers the surface features only. The traditional POS tagging rules ignores the semantic meaning and relatedness of terms, they defines the lexical match of terms.

Usually a search query can represent any object or events in a short way by using keywords such as name, place name, abbreviation, nick name, or events like Second World War. Terms may relate to more than one concept, where concepts are the glue that holds a person's mental world. These terms may be an instance of a word, or a concept in which, it can conceptualize to many related objects or merely attributes.

The proposed framework identifies the intent of the user query at an earlier stage. The semantic meanings of a query word is different for different people, but it may have some common meanings also. The proposed system measure the co-occurrence count of a typed term by considering semantic relatedness and similarity between terms. The concepts for a term have taken as intend nodes in the knowledge base. From the term graph the paths with highest affinity scores are traced out. The nodes with high concept density on the path graph are dynamically noted. The count of cycles in the term graph shows the highest popularity of the term, but result set may be large. A fuzzy membership measuring function has derived to determine the uncertainty among terms. The terms with lower membership value can remove from the path list. The inference clustering has arranged the concepts, random keywords and attributes on a multi-dimensional spatial graph. A co-occurrence based filtering has done to pull up more important and popular result set. The experiments carried out with the proposed system on synthetic and real data sets have listed in section named Experimental Analysis.

This paper having the following contributions:

1. The proposed keyword diversification framework considers the semantic relatedness of the words in a search query and can answer even for vague and short keyword.
2. The inference clustering allows adjusting the search space based on the concept density.
3. The dimensionality reduction method reduces the dimension of feature set based on the normalized projection error method to a considerably minimum level.
4. The XML data model is utilized for precise representation of unstructured data.

MOTIVATION

The traditional methods for text understanding will not deal with the type ambiguity problem in topic modeling. When to project the reduced feature set to a projection plane, the clusters in the joint data are overlapping [1]. Even with transformed features the overlapping will not show a considerable discrimination [2], [3]. The Eigen values of PCA inflated the component loading. The use of combined features with "n" original features improves the situation. Linear Discriminant Analysis (LDA) provides better separation among the classes, and provides better dimension reduction but can't always offer better reduced result sets. The knowledge base containing various forms of data, an inference clustering should be more scalable.

RELATED WORKS

INTENT-BASED DIVERSIFICATION BY CONCEPTUALIZATION

NER locates named entities in a text or tweets and classifies them in a predefined category such as persons, organization, and locations etc. [4], [5]. Topic modeling is based on statistical model. It is used for discovering abstract topics from a collection of documents; it is widely used for text modeling, collaborative filtering and text classification. Latent Dirichlet allocation model (LDA) is the first graphical model for topic modeling [6]. In Fuzzy bag of word model a text is represented as the multi set of its words.

The act of determining the identity of entities in a text with their corresponding entities in a knowledge base is known as entity linking [7]. Conceptualization is the process of select a subset, which contains the objects, concepts and entities of some specified interest for some purpose. Ontology is the explicit specification of conceptualization. Naming and Defining the type, properties and interrelationship of entities in particular domain of discourse is known as ontology. It is the compartmentalization of variables and attributes used for a computation and establishes the relationships among them.

The average number of attributes, features and binary relations issued from a given concept node is known as knowledge density. If semantic coherence is considering, the traditional Longest Cover Method for text segmentation is not suitable. The state-of-the art text segmentation methods will not consider the semantic relationships among various instances and concepts. The proper context of the term in a sentence can identify only by the thorough knowledge about the terms in text segmentation, part of speech tagging, and concept labeling process. This will help to harvest the proper semantic of a word [8]. The lexical semantic relationships among terms can be extracted from a knowledge base, web corpus, probabilistic network, or a probabilistic database [9], [10].

TEXT SEGMENTATION

Text segmentation is the process of splitting a sentence or a phrase of a search query to a set of suitable processing units [11]. The statistical and vocabulary based approach are using for text segmentation. Segments serving the same intention can identify earlier and can group to an intention clusters [12]. There are two types of similarity measures for short text, symmetric and asymmetric [13]. The symmetric approach such as cosine similarity with TF-IDF is widely used for various Natural Language Processing (NLP), Information Retrieval (IR) and tasks. The asymmetric methods are used for various modeling techniques.

ENTITY LINKING

Entity linking is the act of linking the entity mentioned in the text with their corresponding entities in the knowledge base. Candidate entity generation, Candidate entity ranking, and Unlinkable mention prediction are the various phases of entity linking [14]. In Collective entity linking various entities in the documents is considered [15].

SEMANTIC REPRESENTATION WITH FUZZY MODEL

Topic models such as Probabilistic latent semantic analysis (PLSA) [16], [17], Pachinko allocation model (PAM) [18], Latent Dirichlet Allocation (LDA) [6] are used to detect instructive structures in data. The dirichlet is sampled for each document and a multinomial topic node is sampled repeatedly with in the document.

A fuzzy set is a various possible values in a domain. A fuzzy model represents the degree of the semantic association with the domain of the fuzzy set and also represent the uncertainty in the degree using membership function [19], [20]. In Fuzzy Bag-of-Words (FBoW) model [21] the document mapping is achieved by semantic correlation among words quantified by cosine similarity. Documents are represented by using numerical vectors. Pattern mining with weak-wildcard can matches any character in an alphabetic subset [22], [23]. A non linear regression technique that discovers a coherence based separation from highly noisy matches is discussed in [24].

DIMENSIONALITY REDUCTION

Representing documents in low dimensional space can help understanding of relations between documents and the topics they cover [17]. The Normalized Projection Error (NPE) can be used for reducing the feature set and to consider the semi features [1].

PROBLEM STATEMENT

PROBLEM DEFINITION AND FRAMEWORK OVERVIEW

Table 1. The symbols and notations used in the paper.

Notation	Definition
N	Number of concepts in the Database
M	The selected concepts where m<=n
X	Query term
Y	Concept
$CD(x_i,y_j)$	Concept density distribution function
ASi	Affinity score
CCi	Co-occurrence count
Y	Concept density
P(y)	Probability mass function for concept density

Given a text “s” written in a natural language, list out all the cycles for “s” with a given concept in the concept graph and the maximum valued paths are noted. A vector field has created with axes as concepts, terms, instances and attributes.

A concept vector c=p1i+p2j+p3k has defined with i, j, k are unit vectors, a triangle connects the vector to its projection (p1, p2, 0) in the x-y plane. The concept selection can adjust by a threshold value based on the affinity scores. The concept vector is arrayed on the basis of co-occurrence count of instance, concept and attributes.

The maximum number of cycles in the graph and maximum valued paths are listed, the dynamic path detect algorithm will list out more semantically related paths of concepts.

The concept density function has defined as

$$CD(x_i,y_j) = 1\sum^{n-m}(ASi)*1\sum^{n-m}(CCi) \quad (1)$$

The probability mass function for concept density shows which concept should be the more ranked one to the result set.

$$P(y) = (1-p)^{y-1}.p \quad (2)$$

where y is the concept density. The probability mass function for concept density shows which concept should be the more ranked one in the result set.

METHODOLOGY

This section explains the methods followed to design a frame work for keyword search diversification. The Text Segmentation phase has segmented the search query to distinct terms. A synthetic database containing various information has maintained for experimental observations. The data stored in a table of three fields' named term, concept and instance. It helped to design an entity linking mapping (document mapping) for terms for the proposed frame work. The framework can identify the concepts related to the text and can construct a Concept Network. The Concept Density function can easily measure the importance of a term in the given knowledge base. The next step is the link score calculation, we have calculated the affinity score of the link based on the semantic similarity and semantic relatedness by Equation (1). This will represent the strength of semantic relations among concept nodes and their attributes.

A path trace algorithm is used to list out all the cycles in the desired portions of the semantic network. The obtained affinity score on the previous step helped us to create a three dimensional pyramid of clusters based on the values of concept, terms and attributes. This made the frame work to produce huge result set. A dimensionality reduction process has conducted to reduce the result set. Finally the Co-occurrence based filtering is used to pull up the more frequent and important results.

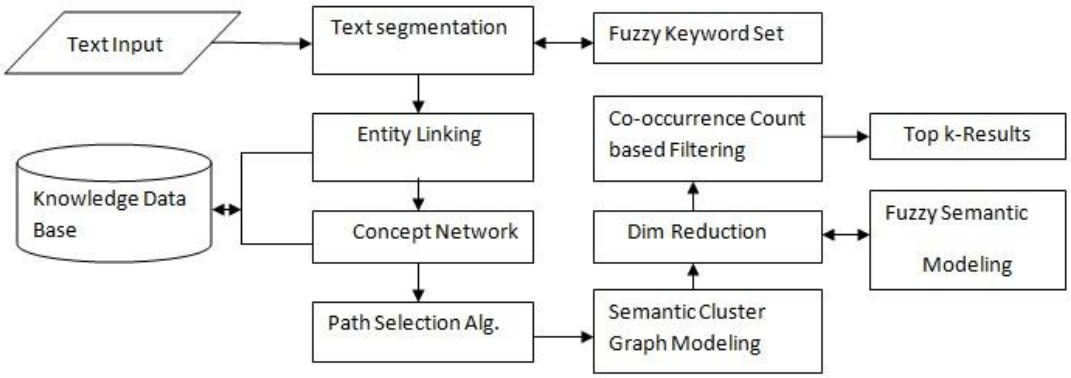


Figure 1: Conceptualization Framework.

TERM EXTRACTION

In this phase the given sentence has segmented in to different processing units. For example, the sentence " california hotel" has segmented into two words "california" and "hotel". In the given synthetic data set hotel is a concept. The term "california" is a named entity, it may have multiple names and the names may have related to several different other named entities, in order to preserve the semantic meaning of a named entity, it could be related to all the concepts in the universal domain for that named entity.

ENTITY LINKING

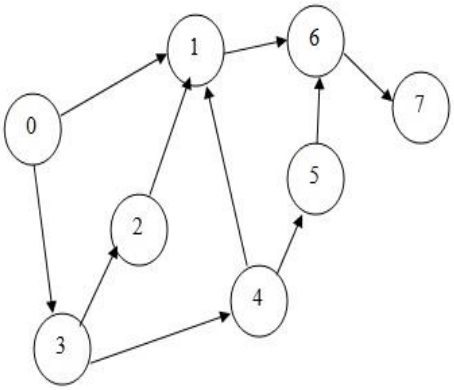
The entity linking procedure has used to leverage the information such as the context of the typed term, and the entity information from the knowledge base to link the named entity with the candidate entities. The proposed method fetches out the concepts in which the given term is related.

CONCEPTUALIZATION

The framework considers the popularity of the terms for conceptualization process to proceed, the popularity it measures based on the criteria such as occurrence of the term in multiple places, the abstract idea of the terms, the co-occurrences of the terms in the knowledge base. The proposed framework easily detects the concepts and instances related to the given term and produce a result set.

The produced result set may be large in numbers, and it may loose the semantic importance of a term, so the proposed work calculates the affinity scores of the concepts in the virtual term graph. As depicted in the

Figure2 the concepts and instances are the nodes in the term graph and concepts are the intent nodes since it has an abstract idea associated with the given term.



The Obtained paths are:

Paths from 3to 1

3 2 1 3 4 1
The Path List

3 2 1 3 4 1

Figure 2a: Term Graph. **Figure 2b: Path table for concept node 3 to node 1.**

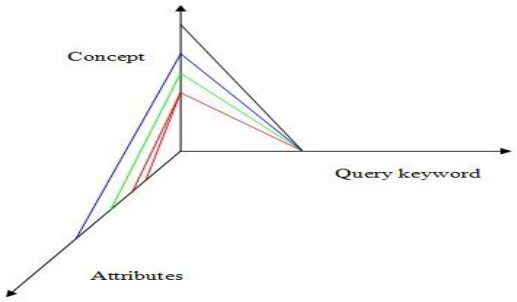
In the term graph the system is identified the cycles formed, and calculated the path length. The path table contains multiple numbers of paths, the paths with highest affinity scores are semantically similar, and it may in contact with more number of concepts. The concept density function (CDF) defines the probability of a path being the more semantically related path for the given data set. The system has used a threshold based voting scheme for concept labeling. The semantically related concepts may be coherent, the more frequently co-occurring terms show high co-occurrence count. The system achieved this by a cross reference with the concept clusters.

FUZZY INFERENCE CLUSTERING

The fuzzy inference clustering is used to shows the interdependence among concept clusters, this is more scalable to shows the inference in a knowledge base. Fig shows the fuzzy inference clustering by taking concepts, terms, and sub categories of concepts and attributes on a multi-dimensional spatial graph. The Figure3 shows a hierarchical fuzzy inference clustering model. The concept cluster vector has arrayed the selected concepts for the typed term. The topics and attributes are chosen by inferring with the concepts in hand.

From Figure 3, For a query keyword “hotel california”, california is the term, and in the synthetic data base it is associated with the concepts {state, song, music, hotel, etc.}. All these concepts are linked to many topics and attributes. The affinity scores for the above concepts shows, (hotel,.58), (state,.43), (song,.51), (music,.42). This shows the term “California” is more related to the concept “hotel”. In concept labeling a threshold value is using to set an outlier for the affinity score and can select top most concepts.

The co-occurrence based filtering is used to list out the most popular search results; Figure 4 illustrates the top most search results.



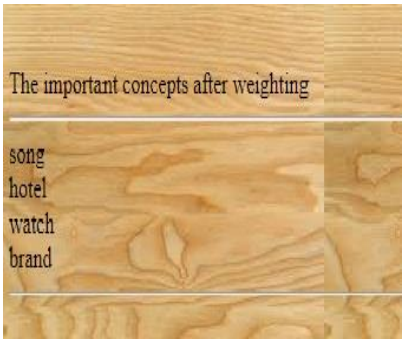
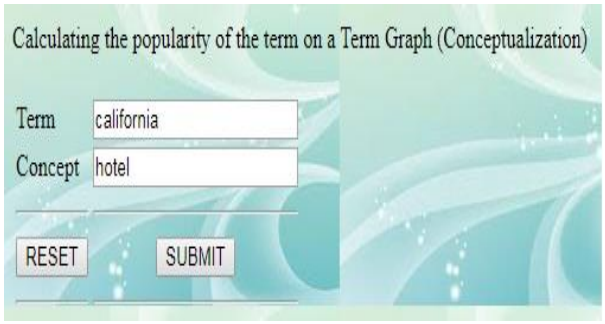


Figure 3: Concept based clustering. **Figure 4: Weighted voting.**

EXPERIMENTAL ANALYSIS

The document mapping of the query term "california" lists out all the concepts related to the typed term such as state, song, music etc. and instances such as troop, books, lyrics, New York, Karnataka, Bangalore, book in the synthetic data set. The list in Figure 5b is not the complete list. All these objects are scattered over the database. The Figure 5a and 5b show the first observation with the proposed system for a query keyword "hotel california". The term california is more prone to be associated to something more than the US state, California.



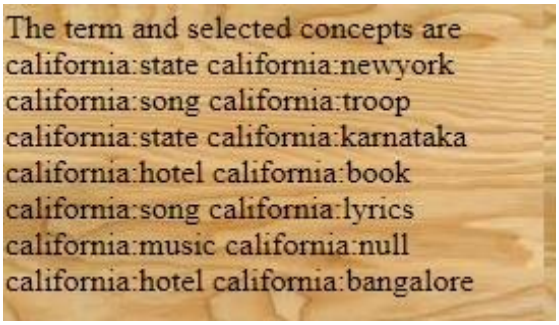


Figure 5a: Conceptualization UI. **Figure 5b: The entity table for the term California.**

The proposed system has been implemented on synthetic data set and a portion of Probase data set. The observations are listed out in the following Table 2.

Table 2. The experimental analysis of the system.

Proposed System		Precision	Semantic similarity	Concept density mass function
Query	Text	0.67	.5042	0.67

The similar values for precision and probability mass function shows the most recent occurrence of the success rate is more in the proposed system.

CONCLUSION

Data mining is the process in which the logical inference of data is performing along with its various attributes, features, and their associations [25], [26]. The proposed system initially segmented the search query, and for the selected typed term it identified the concepts and instance by an entity linking approach. The path detect process calculated the affinity scores of the term graph, and a multi-dimensional search space has formed with concepts, terms and topics. Inference clustering is used to cluster the result set. Concept labeling has done to vote over the cluster model, and a co-occurrence based filtering is used to list out the most popular results.

REFERENCES

I. S. Murthy, CA, "Bridging Feature Selection and Extraction: Compound Feature Generation," IEEE Transactions on Knowledge and Data Engineering, vol. 29, no. 4, pp. 757–770, 2017.
II. I. T. Jolliffe, "Principal Component Analysis and Factor Analysis," pp. 115–128, 1986.
III. H. Abdi and L. J. Williams, "Principal Component Analysis," Wiley interdisciplinary reviews: computational statistics, vol. 2, no. 4, pp. 433–459, 2010.
IV. W. Hua, Z. Wang, H. Wang, K. Zheng, and X. Zhou, "Understand Short Texts by Harvesting and Analyzing Semantic Knowledge," IEEE Transactions on Knowledge and Data Engineering, vol. 29, pp. 499–512,2017.3

V. C. Li, A. Sun, J. Weng, and Q. He, "Tweet Segmentation and Its Application to Named Entity Recognition," IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 2, pp. 558–570, 2015.
VI. D. M. Blei, A. Y. Ng, and M. I. Jordan, Latent Dirichlet Allocation, 2002.
VII. W. Shen, J. Wang, and J. Han, "Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions," IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 2, pp. 443–460, 2015.
VIII. W. Hua, Z. Wang, H. Wang, K. Zheng, and X. Zhou, "Short Text Understanding through Lexical-Semantic Analysis," ICDE, pp. 495–506, 2015.
IX. Y. Song, H. Wang, Z. Wang, H. Li, and W. Chen, "Short Text Conceptualization using a Probabilistic Knowledgebase."
X. X. Liu, Y. Song, S. Liu, and H. Wang, "Automatic Taxonomy Construction from Keywords," Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 1433–1441,2012.
XI. J.T. Chien and C.H. Chueh, "Topic-based Hierarchical Segmentation," IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 1, pp. 55–66, 2012.
XII. D. Papadimitriou, G. Koutrika, Y. Velegrakis, and J. Mylopoulos, "Finding Related Forum Posts through Content Similarity over Intention-based Segmentation," IEEE Transactions on Knowledge and Data Engineering, vol. 29, no. 9, pp. 1860– 1873, 2017.
XIII. J. Jeon, W. B. Croft, and J. H. Lee, "Finding semantically similar questions based on their answers," pp. 617–618, 2005.
XIV. W. Shen, J. Wang, and J. Han, "Entity linking with a knowledge base: Issues, techniques, and solutions," IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 2, pp. 443–460, 2015.
XV. X. Han, L. Sun, and J. Zhao, "Collective Entity Linking in Web Text: A Graph-based Method," pp. 765–774, 2011.
XVI. T. Hofmann, "Probabilistic Latent Semantic Indexing," vol. 51, no. 2, pp.211–218, 2017.
XVII. D. A. Cohn and T. Hofmann, "The Missing Link-a Probabilistic Model of Document Content and Hypertext Connectivity," pp. 430–436, 2001.
XVIII. W. Li and A. McCallum, "Pachinko Allocation: DAG-structured Mixture Models of Topic Correlations," pp. 577–584, 2006.
XIX. T.H. Chang and C.H. Lee, "Subtopic Segmentation for Small Corpus using a Novel Fuzzy Model," IEEE Transactions on Fuzzy Systems, vol. 15, no. 4, pp. 699–709, 2007.
XX. I. J. Chiang, C. C. H. Liu, Y. H. Tsai, and A. Kumar, "Discovering Latent Semantics in Web Documents using Fuzzy Clustering," IEEE Transactions on Fuzzy Systems, vol. 23, no. 6, pp. 2122–2134, 2015.
XXI. R. Zhao and K. Mao, "Fuzzy Bag-of-Words Model for Document Representation," IEEE Transactions on Fuzzy Systems, 2017.
XXII. C.D. Tan, F. Min, M. Wang, H.R. Zhang, and Z.H. Zhang, "Discovering Patterns With Weak-Wildcard Gaps," IEEE Access, vol. 4, pp. 4922–4932, 2016.
XXIII. J. Li, Q. Wang, C. Wang, N. Cao, K. Ren, and W. Lou, "Fuzzy keyword search over encrypted data in cloud computing," INFOCOM, 2010 Proceedings IEEE, pp. 1–5, 2010.
XXIV. W.Y. Lin, F. Wang, M. M. Cheng, S.-K. Yeung, P. H. Torr, M. N. Do, and J. Lu, "Code: Coherence based Deecision Boundaries for Feature Correspondence," IEEE transactions on pattern analysis and machine intelligence, 2017.
XXV. S. P. K.R.Venugopal, Champa H.N., "A Survey on Intent-based Diversification for Fuzzy Keyword Search over XML Data," IJCSIT, vol. 8, no. 6, pp. 612–618, 2017.
XXVI. B. Xue, M. Zhang, W. N. Browne, and X. Yao, "A Survey on Evolutionary Computation Approaches to Feature Selection," IEEE Transactions on Evolutionary Computation, vol. 20, no. 4, pp. 606–626, 2016.