ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

A Survey on Similarity Detection in Text

Priyanka R.Patil[1], Shital A.Patil[2]

PG Student, Dept. of Computer Engineering, SSBT'S COET, Bambhori, North Maharashtra University, Jalgaon, India[1]

Assistant Professor, Dept. of Computer Engineering, SSBT's COET, Bambhori, North Maharashtra University, Jalgaon, India[2]

ABSTRACT: SimilarityView is an application for visually comparing and exploring multiple models of text corpora. SimilarityView uses multiple linked views to visually analyze both the conceptual content and the document relationships in models generated using different algorithms. Existing system for Filtering the text is difficult in the basic text classification. This system is not finding synonyms and matching graph it only work on single word calculation and normal text matching. Friendbook discovers life styles of users from user-centric sensor data, measures the similarity of life styles between users, and recommends friends to users if their life styles have high similarity. Inspired by text mining, to model a user's daily life as life documents, from which his/her life styles are extracted by using the Latent Dirichlet Allocation algorithm. The proposed method to finding the proper matching graph use Latent dirichlet Allocation (LDA) algorithm. A method like- text mining method came into picture to solve the problem of automatically checking the paragraph semantically.This approach uses Term Frequency- Inverse Document Frequency (TF-IDF) and Latent Semantic Indexing (LSI) to semantically find SimilarityView.

KEYWORDS: Semantic similarity,Document Similarity Graphs, Text mining.

I. INTRODUCTION

In social networking services recommend friends to users based on their social graphs, which may not be the most appropriate to reflect a user's preferences on friend selection in real life. In this paper, we present Friendbook, a novel semantic-based friend recommendation system for social networks, which recommends friends to users based on their life styles instead of social graphs. By taking advantage of sensor-rich Smart phones, Friendbook discovers life styles of users from user-centric sensor data, measures the similarity of life styles between users, and recommends friends to users if their life styles have high similarity.

Inspired by text mining, we model a user's daily life as life documents, from which his/her life styles are extracted by using the Latent Dirichlet Allocation algorithm. Further propose a similarity metric to measure the similarity of life styles between users, and calculate users' impact in terms of life styles with a friend-matching graph. Upon receiving a request, Friendbook returns a list of people with highest recommendation scores to the query user. Finally Friendbook integrates a feedback mechanism to further improve the recommendation accuracy. We have implemented Friendbook on the Android-based smart phones, and evaluated its performance on both small scale experiments and large-scale simulations.

The importance of contextual information has been recognized by researchers and practitioners in many disciplines including Ecommerce, personalized IR, ubiquitous and mobile computing, data mining, marketing and management. There are many existing e-commerce websites which have implemented recommendation systems successfully. We will discuss few website in our coming section that provides recommendation. Items are suggested by looking at the behavior of like-minded-users. Groups are formed of such users, and items preferred by such groups are recommended to the user, whose liking and behavior is similar to the group. In our model we have incorporated user preferences obtained from Social Networking Site. Social Networking sites are used intensively from last decade. According to the current survey, Social Networking sites have the largest data set of users. Each social networking site notes/records each and every activity of user (like: what user likes? what user is doing? what is user's hobby? Etc).

The rest of the paper is organized as follows: In Section II we review the literature survey. Various steps of plagiarism are described in Section III. In Section IV we present the proposed mechanism. We conclude the paper in Section V.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

II. LITERATURE SURVEY

To assess how well existing methods model human semantic memory, Griffiths et al. [3] compare generative probabilistic topic models with models of semantic spaces.They are concerned with a model's ability to extract the gist of a word sequence in order to disambiguate terms that have different meanings in different contexts. This is also related to predicting related concepts. LSA and LDA are used as instances of these approaches and compared in word association tasks.

In contrast, our work focuses on comparing the impact that model differences have on visual analytics applications, using visualization to do the comparison.Collins et al. [4] combine tag clouds with parallel coordinates to form Parallel Tag Clouds, an approach for comparatively visualizing differentiating words within different dimensions of a text corpus.

Word lists are alphabetical, with word size scaled according to word weight. Similar to parallel coordinates, matching terms are connected across columns. Although the similar goals in comparing term lists, to feel that this approach of sorting terms by weight, combined with scaling text luminance by weight, provides a clear comparison of the relative significance of terms across concepts and topics. This avoids the layout complications and potential overlaps encountered when words are drawn at vastly different scales. The pre-processing operations are carried out by three components; the tokenizer, stop-word removal component and case folding component.

III. PLAGIARISM DETECTION ALGORITHM (PDA)

The main plagiarism process consists of three steps, as follows

- Text document collection
- Text document preprocessing
- Text document encoding

The main plagiarism process consists of four steps, as follows:

A. *TEXT DOCUMENT COLLECTION:*

The existing research papers are stored in the text format, within the database.

B. *TEXT DOCUMENT PRE-PROCESSING:*

The contents of papers are usually non-structured. The pre-processing analyzes, extracts, and identifies the keywords in the full text of the papers and tokenizes them. Here, a further reduction in the vocabulary size is achieved, through the removal of frequently occurring words referred as stop-words, via- stop file. This is called as filtering phase of removal of stop word.

C. *TEXT DOCUMENT ENCODING:*

On filtering text documents they are converted into a feature vector. This step uses TF-IDF algorithm. Eachtoken is assigned a weight, in terms of frequency (TF), taking into consideration a single research paper. IDFconsiders all the papers, scattered in the database and calculates the inverse frequency of the token appeared inall research papers. So, TF is a local weighting function,while IDF is global weighting function.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

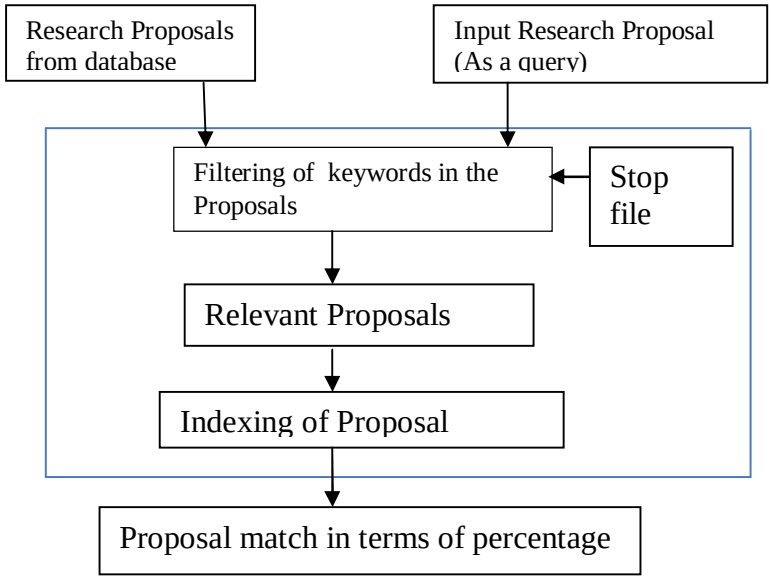

Fig: System Architecture

IV. PROPOSED ALGORITHM

The system finally outputs the Best Matching Unit (BMU). The system provides Best 5 matched papers with respect to the input research paper, in the descending order, with the ordered best matched paper. After the research papers are submitted by the end-users, the papers in provided discipline are checked using the text-mining technique.

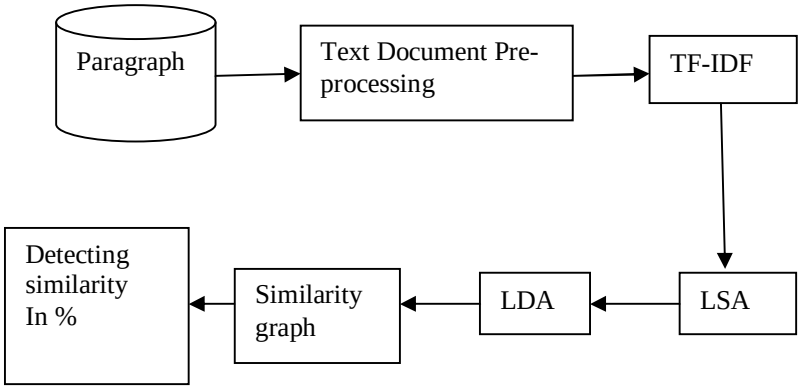

Fig: Text Mining

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

A. TF-IDF:

TF-IDF encoding describes a weighted method based on inverse document frequency (IDF) [7] combined with the term frequency (TF) to produce the feature v, such that vi = tfi*log (N \dfi) The weights are assigned using above formula. Here, N is the total number of papers in the discipline, tfi is the term frequency of the feature word wi and dfi is the number of papers containing the word wi. TF increases the weight of term and IDF decreases weight of term. The Term-Document matrix is created in this stepas shown in fig.

B. LATENT SEMANTIC ANALYSIS:

LSA computes a truncated SVD of a term-documentmatrix [5], i.e., the collection of weighted term vectorsassociated with the documents in a corpus of text. Morespecifically, the *k*-dimensional LSA model of a term-document matrix , A □ Rm ×n, is its rank-k SVD,

$$A_k = U_k \Sigma_k V_k^T ,$$

where$Uk \in \mathbb{R}m \times k$, $\Sigma k \in \mathbb{R}k \times k$, $Vk \in \mathbb{R}n \times k$ contain the *k*leading left singular vectors, singular values, and rightsingular vectors, respectively. The *k* latent features, orconcepts, are linear combinations of the original terms,with weights specified in *Uk*. Documents are modelled as vectors in concept space ,with coordinates specified in Vk.

C. LATENT DIRICHLET ALLOCATION:

LDA is a hierarchical probabilistic generative approachthat models a collection of documents by topics, i.e.probability distributions over a vocabulary [2]. Given avocabulary of *W* distinct words, a number of topics *K*, twosmoothing parameters α and β, and a prior distributionover document lengths (typically Poisson) – thisgenerative model creates random documents whosecontents are a mixture of topics. In order to use LDA tomodel the topics in an existing corpus, the parameters ofthe generative model must be learned from the data.Specifically, for a corpus containing *D* documents wewant to learn φ, the $K \times W$ matrix of topics, and θ, the $D \times K$matrix of topic weights for each document. The remainingparameters α,β and *K* are specified by the user. For theLDA models used in this paper, parameter fitting isperformed using collapsed Gibbs sampling [8] to estimateθandφ.

V. CONCLUSION

Using SimilarityView, we find that LSA concepts providegood summarizations over broad groups of documents,while LDA topics are focused on smaller groups. LDA'slimited document groups and its probabilistic mechanismfor determining a topic's top terms support better labellingfor document clusters than LSA concepts, but thedocument relationships defined by the LSA model do notinclude extraneous connections between disparate topics identified by LDA in our example.

REFERENCES

1. Zhibo Wang., Student Member, IEEE, Jilong Liao., Qing Cao., Member, IEEE, Hairong Qi., Senior Member, IEEE, and Zhi Wang., Member, IEEE., "Friendbook: A Semantic-based Friend Recommendation System for Social Networks", IEEE Transactions on Mobile Computing.
2. G. R. Arce.,"Nonlinear Signal Processing: A Statistical Approach". John Wiley & Sons, 2005.
3. J. Biagioni, T.,Gerlich., T. Merrifield., and J. Eriksson.,"Easy Tracker: Automatic Transit Tracking, Mapping, and Arrival Time Prediction Using Smart phones", In Proc. of Sen Sys, pages 68–81, 2011.
4. L. Bian and H. Holtzman. "Online friend recommendation through personality matching and collaborative filtering", In Proc. Of UBICOMM, pages 230–235, 2011.
5. C. M. Bishop.," Patternrecognition and machine learning" Springer New York, 2006.
6. D. M. Blei., A. Y. Ng., and M. I. Jordan., " Latent Dirichlet Allocation. Journal of Machine Learning Research", 3:993– 1022, 2003.
7. N. Eagle., and A. S. Pentland., "Reality Mining: Sensing Complex Cocial Systems. Personal Ubiquitous Computing", 10(4):255–268, March 2006.
8. K. Farrahi., and D. Gatica-Perez.," Probabilistic mining of sociogeographic routines from mobile phone data. Selected Topics in Signal Processing", IEEE Journal of, 4(4):746–755, 2010.
9. Kwon., and Kim., "Similarity Recognition techniques", In Proc. Of UBICOMM., pages230235, 2013.
10. S.C. Deerwester.,S. T. Dumai., T. K. Landaueer.,G.W. Furnas., and R. A. Harshman., "Indexing by Latenet Semanatic Analysis"., JASIS, vol. 41, no.6, pp.391-407, 2009.
11. M. W. Berry., S. T.Dumais., and G.W.O'Brien., "Using linear algebra for intelligentinformation retrival", SIAM Review, vol.37, no.4,pp.573-595, 2010.
12. P.Over., and J. Yen., "A unified toolkit for information and scientific visualizing", in Proc.Visualization and data analysis, vol.7243. SPIE, 2009.
13. G.S.Davidson., B. Hendrickson., "Knowledge mining with vxinsight: Discovery through interaction", Journal of Intelligent information Systems, vol. 11, no. 3, pp.259-285,2012.