International Journal of Soft Computing 8 (3): 149-153, 2013
ISSN: 1816-9503

Web Page Clustering Based on Novel Latent Semantic Approach

P. Manimaran and K. Duraiswamy
Department of Computer Science and Engineering,
K.S. Rangasamy College of Technology, Namakkal-Dt., Tamilnadu, India

Abstract: Clustering algorithms are usually based on the Bag-of-Words (BOW) approach. A tarnished hindrance of the BOW prototypical is that it ignores the semantic relationship among words. As a result, if two documents use different collections of core words to represent the same topic, they may be assigned to different clusters even though the core words they use are probably synonyms or semantically associated in other form and other disadvantage of conventional web page clustering technique is often utilized to reveal the functional similarity of WebPages. Tagging can be beneficial to improve the clustering performance. Several efforts have been made to explore social tagging for clustering. But there is some drawbacks of tagging web based clustering. All the existing approaches exploiting tag information for web page clustering assume that all the WebPages are tagged which is a somewhat restrictive assumption. In a more realistic setting, one can only expect that the tags will be available for only a small number of WebPages. Researchers propose a new web page grouping approach based on Probabilistic Latent Semantic Analysis (PLSA) Model. An iterative set of rules based on maximum likelihood principle is employed to overcome the aforementioned computational shortcoming.

Key words: Probabilistic latent semantic analysis, singular value decomposition, term-frequency, web page clustering, India

INTRODUCTION

The main goal is to present a multiparty probabilistic model of document to find the discriminant feature of WebPages and find the actual relation between word and page. Huge repositories of textual data have become available to a large public. Today, it is one of the great challenges in the information sciences to develop intelligent interfaces for human machine interaction which support computer users in their quest for relevant information. Although, the use of elaborate ergonomic elements like computer graphics and visualization has proven to be extremely fruitful to facilitate and enhance information access, progress on the more fundamental question of machine intelligence is ultimately necessary to ensure substantial progress on this issue. In order for computers to interact more naturally with humans, one has to deal with the potential inconsistency, impreciseness or even vagueness of user requests and has to recognize the difference between what a user's might say or do and what she or he actually meant or intended. One typical scenario of human machine interaction in information retrieval is by natural language queries: the user formulates a request, e.g. by providing a number of keywords or some free-form text and expects the system to return the relevant data in some amenable representation, e.g. in form of a ranked list of relevant documents. Many retrieval methods are based on simple word matching strategies to determine the rank of relevance of a document with respect to a query. It is well known that literal term matching has severe drawbacks, mainly due to the ambivalence of words and their unavoidable lack of precision as well as due to personal style and individual differences in word usage. Latent Semantic Analysis (LSA) (Hofmann, 1999a, b) is an approach to automatic indexing and information retrieval that attempts to overcome these problems by mapping documents as well as terms to a representation in the so called latent semantic space. LSA usually takes the (high dimensional) vector space representation of documents based on term frequencies as a starting point and applies a dimension reducing linear projection. The specific form of this mapping is determined by a given document collection and is based on a Singular Value Decomposition (SVD) of the corresponding term/document matrix. The general claim is that similarities between documents or between documents and queries can be more reliably estimated in the reduced latent space

Corresponding Author: P. Manimaran, Department of Computer Science and Engineering, K.S. Rangasamy College of Technology, Namakkal-Dt., Tamilnadu, India

representation than in the original representation. The rationale is that documents which share frequently co-occurring terms will have a similar representation in the latent space even if they have no terms in common. LSA thus performs some sort of noise reduction and has the potential benefit to detect synonyms as well as words that refer to the same topic. In many applications, this has proven to result in more robust word processing. Although, LSA has been applied with remarkable success in different domains including automatic indexing (Latent Semantic Indexing, LSI) (Hofmann, 1999a, b; Dumais, 1995), it has a number of deficits, mainly due to its unsatisfactory statistical foundation. The primary goal of this study is to present a novel approach to LSA and factor analysis called Probabilistic Latent Semantic Analysis (PLSA) that has a solid statistical foundation since it is based on the likelihood principle and defines a proper generative model of the data (Dempster *et al*., 1977). This implies in particular that standard techniques from statistics can be applied for questions like model fitting, model combination and complexity control. In addition, the factor representation obtained by PLSA allows to deal with polysemous words and to explicitly distinguish between different meanings and different types of word usage (Hofmann, 1999a, b).

WEB PAGE CLUSTERING

Traditional webpage clustering typically uses only the page content information usually, just the page text in an appropriate feature vector representation such as Bag of Words, Term-Frequency/Inverse-Document-Frequency, etc. and then applies standard clustering algorithms, e.g., K-means algorithm, spectral clustering, etc. Traditional webpage clustering has some drawbacks that reducing the speed of searching make navigation problem and quality of clustering is not good. Researchers are facing an ever increasing volume of text documents. The abundant texts flowing over the Internet, huge collections of documents in digital libraries and repositories and digitized personal information such as blog articles and emails are piling up quickly every day. These have brought challenges for the effective and efficient organization of data. These have brought challenges for the effective and efficient organization of text documents. Clustering in general is an important and useful technique that automatically organizes a collection with a substantial number of data objects into a much smaller number of coherent groups. In the particular scenario of text documents, clustering has proven to be an effective approach for quite some time and an interesting research problem as well. It is becoming even more interesting and demanding with the development of the World Wide Web.

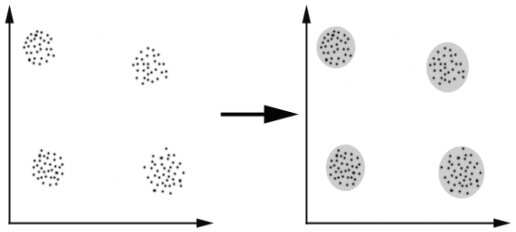


Fig. 1: Graphical example of clustering

What is clustering?: Clustering can be considered the most important unsupervised learning problem so, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data. A loose definition of clustering could be the process of organizing objects into groups whose members are similar in some way. A cluster is therefore a collection of objects which are similar between them and are dissimilar to the objects belonging to other clusters. Researchers can show this with a simple graphical example (Fig. 1).

In this case, researchers easily identify the 4 clusters into which the data can be divided; the similarity criterion is distance: two or more objects belong to the same cluster if they are close according to a given distance. This is called distance-based clustering. Another kind of clustering is conceptual clustering: two or more objects belong to the same cluster if this one defines a concept common to all that objects. In other words, objects are grouped according to their fit to descriptive concepts not according to simple similarity measures.

PROPOSED SYSTEM

In this study, researchers suggest some alternatives which will make it possible to exploit tag information even when the tag information in available for only a small number of WebPages. The drawback of above mentioned traditional web page clustering solved by web page tagging: How can tagging data be used to improve web document clustering? This is part of a major trend in information retrieval to make more and better use of the increasingly prevalent user-provided data. Social bookmarking web sites such as del.icio.us and Stumble Upon enable users to tag any web page with short free-form text strings, collecting hundreds of thousands of key word annotations per day. The set of tags applied to a document is an explicit set of key words that users have found appropriate for categorizing that document within their Own Filing System. Thus, tags promise a uniquely well suited source of information on the

similarity between web documents. While others have argued that tags hold promise for ranked retrieval including at least one approach that uses clustering. High quality clustering based on user-contributed tags has the potential to improve all of the previously stated applications of the cluster hypothesis from user interfaces to topic-driven language models to increasing diversity of results (Gildea and Hofmann, 1999).

Therefore, such user generated content can provide useful information in various form such as meta-data or in more explicit ways such as tags. User specified tags in particular have proven to be extremely effective in browsing, organizing and indexing of web page. Various social bookmarking websites such as stumble upon and delicious allow users to tag WebPages with key words or short text snippets that can provide a description of the WebPages. Users can collaboratively tag WebPages and this has made organizing, sharing, navigating and retrieving web content much easier than ever before. The aim to exploit the tag information for a webmining task, namely webpage clustering. Since, user provided tags can often be very discriminative for WebPages researchers want to exploit them by treating the tag information as an alternate view of the data. Motivated by the success of multi-view learning algorithms in various machine learning tasks, researchers use two views of the data to extract highly discriminative features and perform clustering using these features (Hofmann *et al*., 1999).

The feature extraction amounts to performing clustering in a lower dimensional subspace which is also effective in dealing with the problem of over fitting when researchers only have a small number of documents having a very large number of features. In particular, researchers use a regularized variant of the kernel canonical correlation analysis algorithm (Trivedi *et al*., 2010) to learn this subspace. Has received tremendous attention due to its ability for effectively extracting useful features from heterogeneous or parallel data sources such as images and text or features and labels supervised dimensionality reduction? Therefore, such an approach is expected to be useful for extracting useful features in the case of webpage clustering as well since the data often does have multiple views.

Advantages:

- Tagging has made organizing, sharing, navigating and retrieving web content much easier than ever before
- High quality clustering based on user-contributed tags has the potential to improve all of the previously stated applications of the cluster
- Reduced searching time

PROBABILISTIC LATENT SEMANTIC ANALYSIS

Latent semantic analysis: Latent semantic analysis is to map documents and by symmetry terms to a vector space of reduced dimensionality, the latent semantic space. This mapping is computed by decomposing the term/document matrix N with SVD, $N = X\Sigma Y^t$ where X and Y are orthogonal matrices $X^tX = Y^tY = I$ and the diagonal matrix Σ contains the singular values of N. The LSA approximation of N is computed by thresholding all but the largest K singular values in Σ to zero (= Σ) which is rank K optimal in the sense of the L2-matrix norm as is well-known from linear algebra.

Geometry of the aspect model: Now consider the class-conditional multinomial distributions P (.|c) over the vocabulary in the aspect model which can be represented as points on the M-1 dimensional simplex of all possible multinomial's. Via its convex hull, this set of K points defines a K-1 dimensional sub-simplex. The modeling assumption expressed is that all conditional distributions P(.|d) are approximated by a multinomial represent able as a convex combination of the class-conditionals P(.|c). In this geometrical view, the mixing weights P(c|d) correspond exactly to the coordinates of a document in that sub-simplex. This demonstrates that despite of the discreteness of the latent variables introduced in the aspect model, a continuous latent space is obtained within the space of all multinomial distributions. Since, the dimensionality of the sub-simplex is K-1 as opposed to M-1 for the complete probability simplex, this can also be thought of in terms of dimensionality reduction and the sub-simplex can be identified with a probabilistic latent semantic space (Tipping and Bishop, 1999).

The core of PLSA is a statistical model which has been called a aspect model. The latter is a latent variable model for general co-occurrence data which associates an unobserved class variable $c \epsilon C = \{c1.... ck\}$ with each observation, i.e. with each occurrence of a word $w \epsilon W = \{w1... wm\}$ in a document $d \epsilon D = \{d1... dN\}$. In terms of a generative model it can be defined in the following way:

- Select a document d with probability P(d)
- Pick a latent class c with probability P(c|d)
- Generate a word w with probability P(w|c)

As a result one obtains an observed pair (d, w) while the latent class variable c is discarded (Hofmann, 1999a, b). Translating this process into a joint probability model results in the expression:

$$P(d, w) = P(d)P(w|d) \quad (1)$$

Where:

$$P(w|d) = \sum_{c \in C} P(w|c)P(c|d) \quad (2)$$

One has to sum over the possible choices of z which could have generated the observation. The aspect model is a Statistical Mixture Model which is based on two independence assumptions: first, observation pairs (d, w) are assumed to be generated independently; this essentially corresponds to the 'bag {of {words' approach. Secondly, the conditional independence assumption is made that conditioned on the latent class z, words w are generated independently of the specific document identity d. Given that the number of states is smaller than the number of documents (K ‹‹ N), c acts as a bottleneck variable in predicting w conditioned on d. Notice that in contrast to document clustering models document {specific word distributions P(w|d) are obtained by a convex combination of the aspects or factors P(w|c). Documents are not assigned to clusters, they are characterized by a specific mixture of factors with weights P(c|d). These mixing weights over more modeling power and are conceptually very different from posterior probabilities in clustering models and (unsupervised) Naive Bayes Models. Following the likelihood principle, one determines P(d), P(c|d) and P(w|c) by maximization of the log likelihood function:

$$l = \sum_{d \in D} \sum_{w \in W} n(d,w) \log P(w,d) \quad (3)$$

where, n(d, w) denotes the term frequency, i.e., the number of times w occurred in d. This is a symmetric model P(z|d) with the help of Bayes' rule which results in:

$$P(d|w) = \sum_{c \in C} P(c)P(w|c)P(d|c) \quad (4)$$

is a re-parameterized version of the generative model that described by Eq. 1and 2.

SYSTEM ARCHITECTURE

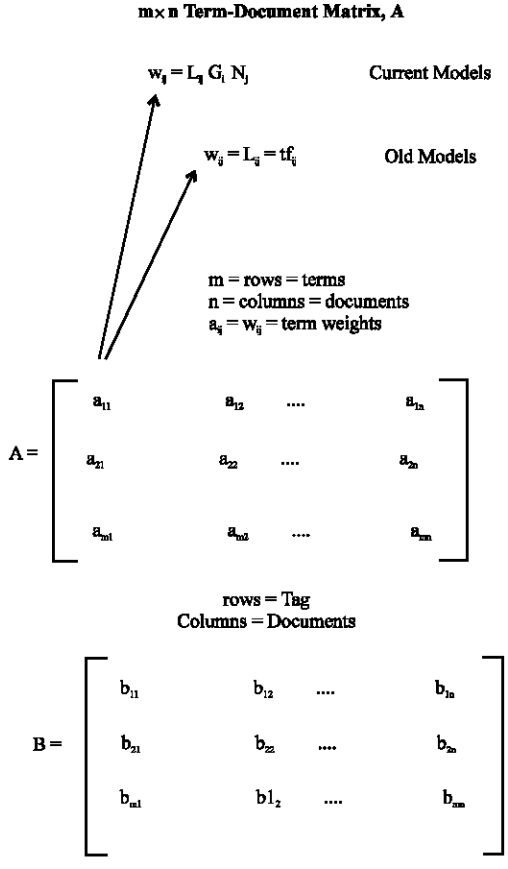
An overview of developed system architecture flow is represented (Fig. 2).

Preprocessing: Retrieving the document snippets from Google and parsing and stemming the results. Delete HTML tags, non-letter characters such as $, % or #. For example, the words: connected, connecting, interconnection should be transformed to the word connect. In third step clear the stop words natural candidates for stop words are articles (e.g., the), prepositions (e.g. for, of) and pronouns (e.g. I, his).

TFIDF calculation: This assigns to term i a weight in document j given by $TF\,IDF_{i,j} = Tf_{i,j} \times IDF_i$

Annotation based probabilistic LSA: Assume that researchers are given two sets of WebPages one set T is tagged and the other set U is non-tagged. Further, |T|‹|U| and N = |T |+|U| is the total number of WebPages. The goal is to obtain a clustering of all N WebPages. Researchers define the following.

A is document-word co-occurrence matrix (bag-of-words representation) of size N×|W| where N is the number of documents (WebPages) in the corpus and |W| is the page-text vocabulary size. Aij denotes the frequency of the word j appearing in document i. Note that the document-word co-occurrence matrix is constructed using both tagged and non-tagged WebPages. B = tag-word co-occurrence matrix (bag-of-words representation) of size |T|×|W| where |T| is the total number of tags in the corpus and |W| is the page-text vocabulary size. B_{ij} denotes the number of times tag i is associated with word j. Having constructed the document-word and word-tag co-occurrence matrices A and B, joint PLSA can be applied using A and B.



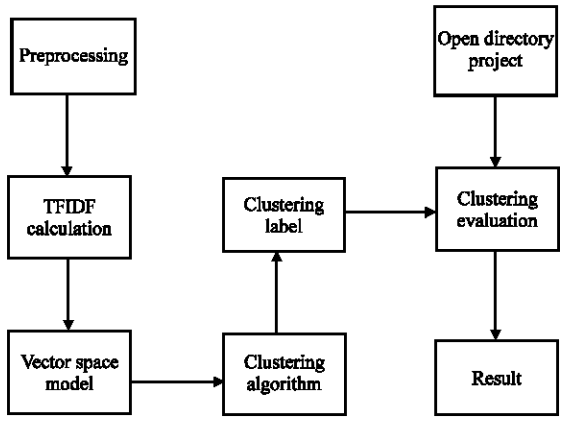


Fig. 2: System architecture diagram

K-means clustering algorithm:

1 Choose cluster centroids to coincide with K randomly selected documents from the document set
2 Assign each document to its closest cluster
3 Recomputed the cluster centroids using the current cluster memberships
4 If reassignment of documents to the new cluster, go to step 2. Typical stopping criteria are: groups formed by subsequent iterations must be

DATASETS

The dataset required for this project is a dataset that consist of user generated data. This is available on delicious social bookmarking website. For evaluation purpose the directory structure that is required is available on the ODP (Open Directory Project) site. The system used the bag-of-words representation for the feature vectors. The approach can however also be applied with other feature representations such as the Term-Frequency/Inverse-Document-Frequency (TF/IDF). A number of techniques have been proposed in the past to improve information retrieval tasks using auxiliary sources information, e.g., anchor text for web search interconnectivity of WebPages captions for image retrieval, etc. Recent works on exploiting social annotations.

CONCLUSION

Probabilistic semantic analysis has important theoretical advantages over standard LSA since it is based on the likelihood principle defines a generative data model and directly minimizes word perplexity. Probabilistic Latent Semantic Analysis (PLSA) that has a solid statistical foundation since, it is based on the likelihood principle and defines a proper generative model of the data. Researchers propose a new web page grouping approach based on Probabilistic Latent Semantic Analysis (PLSA) Model. An iterative algorithm based on maximum likelihood principle is employed to overcome the aforementioned computational shortcoming.

REFERENCES

Dempster, A.P., N.M. Laird and D.B. Rubin, 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc., 39: 1-38.

Dumais, S.T., 1995. Latent semantic indexing Trec-3 report. In: Overview of the Third Text REtrieval Conference (TREC-3), Harman, D. (Ed.). DIANE Publishing, UK, pp: 219-230.

Gildea, D. and T. Hofmann, 1999. Topic-based language models using em. Proceedings of the 6th European Conference on Speech Communication and Technology, September 5-9, 1999, Budapest, Hungary.

Hofmann, T., 1999a. Probabilistic latent semantic analysis. Proceedings of the 15th Conference on Uncertainty in AI, Jul 30-August 1, 1999, Stockholm, Sweden.

Hofmann, T., 1999b. Probabilistic latent semantic indexing. Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 15-19, 1999, Berkeley, CA., USA., pp: 50-57.

Hofmann, T., J. Puzicha and M.I. Jordan, 1999. Unsupervised learning from dyadic data. Advances in Neural Information Processing Systems, International Computer Science Institute.

Tipping, M. and C. Bishop, 1999. Probabilistic principal component analysis. J. Royal Stat. Soc. Series B, 61: 611-622.

Trivedi, A., P. Rai and S.L. DuVall, 2010. Exploiting tag and word correlations for improved webpage clustering. Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents, Toronto, Ontario, Canada, October 30, 2010, ACM, New York, USA., pp: 3-12.