Clustering and Recommendation using WordNet

Justina G. Nadar
P.G Student
Department of Information Technology, Mumbai University
PIIT, New Panvel, India

ABSTRACT
The recommender systems are new type of software tools designed to help users find their way through today's online shops. Due to the increasing number of e-commerce websites, it is necessary to render effective recommendation to the users. Here we present an overview of current recommendation systems and then our proposed system that employs WordNet dictionary in clustering and content based filtration to provide tailored and friendly suggestions to the user. The proposed system is a user-centric approach that analyses the navigation path of the user and clusters the keyword extracted from WordNet to recommend user articles.

Keywords
Recommender systems, WordNet, Clustering, Advanced Kmeans

1. INTRODUCTION
Nowadays due to the large amount of e-commerce websites hosted on the web it is necessary to effectively render services to the user. The rapid expansion of internet has brought a new market for trading. Although it seems to be beneficial to the users, the users are faced with myriad of choices. A user would normally rely on the opinions and advice of friends and family members but unfortunately even they have limited knowledge.

When users browse through a website they are usually looking for items they find interesting. The interestingness can depend on a number of things such as the information a user is looking for, purchased products. A website is the collection of these interesting items. Some website has their own way of recommending items to a user. A website may have a new product or book in the store and recommend every user about the same. Whereas in amazon.com the recommender system recommends a book purchased with the selected item. It is the same for every visitor.

The rest of the paper is organised as follows: Section 2 provides literature survey; Section 3 comparisons of content based filtering and collaborative filtering; Section 4 describes clustering techniques; Section 5 in details about the proposed system and we make our conclusion in Section 6.

2. LITERATURE REVIEW
In this section, we list the past literature review. The various data pre-processing, clustering, categorization and recommendation techniques are cited here.

Christos Bouras [1] used WordNet based approach to cluster the user interests. The sessions obtained from user's navigation path are enriched using WordNet hypernyms and the hypernyms are clustered to provide more generalised keywords for clustering user data.

Zakaria Elberrichi in [2] Using WordNet for Text Categorization, the author describes that the bag of words representation used for text representation is unsatisfactory as it ignores possible relations between terms. The proposed method extracts generic concepts from WordNet for all the terms in the text then combines them with the terms in different ways to form a new representative vector.

In [3] the author introduced Effective Personalization Recommendation Based on Time Framed Navigation Clustering and Association Mining; the effectiveness of the recommendation methods, with and without time-framed user clustering is investigated and compared. The results showed that the recommendation model built with user clustering by time-framed navigation sessions improves the re-commendation services effectively.

Yanjun Li, the author proposed an algorithm Parallel Bisecting K-means and describes that PBKP [4] algorithm fully exploits the data-parallelism of the bisecting k-means algorithm, and adopts a prediction step to balance the workloads of multiple processors to achieve a high speed-up.

Paul Resnick and Neophytos Iacovou [5] discuss about an open architecture for collaborative filtering of Netnews, the author says that collaborative filters help people make choices based on the opinions of other people. GroupLens is a system for collaborative filtering of netnews, to help people find articles they will like in the huge stream of available articles.

The method that is capable of detecting similarities between documents containing semantically similar but not necessarily lexicographically similar terms is described in the paper, Semantic Similarity Methods in WordNet and their Application to Information Retrieval on the Web [6].

Robert Cooley [7] says that, Data preparation for mining World Wide Web shows an important input to the design tasks is the analysis of how a Web site is being used. Web Usage Mining is the application of data mining techniques to usage logs of large Web data repositories in order to produce results that can be used in the design tasks.

Combining two techniques is introduced by Christos Bouras [8] a combination of the algorithms, which achieves co-operation of the categorization and summarization mechanisms, is introduced in order to enhance text labelling through the personalized summaries that are constructed.

The attribute oriented induction approach is described by Y. Fu [9] the access patterns of web users are extracted from web server log files and then organised into sessions which represent episodes of interaction between the web users and the web server. Using attribute oriented induction; the sessions are generalized according to the page hierarchy which organizes pages according to their generalities. The generalized pages are finally clustered.

3. CONTENT BASED FILTERING Vs COLLABORATIVE FILTERING
Content based filtering is also known as cognitive filtering. Content based filtering where words in document are considered as terms and used in analysing a user profile.

Papers [3][10] describe Content based Recommendation Systems in detail.

Collaborative filtering is also known as social filtering; it filters by using information of other people. Wang [3] compares both content and collaborative system in the paper. Collaborative filtering system is based on the idea that people who agreed on the in their evaluation on certain items in the past are likely to agree again in the future. For Example, A person who wants to read books on science fiction will agree with those who have similar interests like him to read about science fiction.

4. CLUSTERING TECHNIQUES

There are various clustering methods that can be used along with recommendation system. Some of the techniques are listed here.

The most popular and widely used clustering technique is the partitioning method. Though many variations have come along it gives good results such techniques have advanced with time. K-means, advanced kmeans, k mediods, bisecting kmeans, and parallel bisecting kmeans are few methods in partitioning method [12].

The other method is the hierarchical method [12]. It is often combined with summarization and categorization mechanisms. The hierarchical method gives more generalised form of words and it is widely used in text retrieval systems. Some examples are binary tree structure, AGNES, DIANA and BIRCH.

Other techniques [12] are Density based models, Grid models, Probability models, Vector space models and Bayesian models.

5. PROPOSED SYSTEM

The proposed system an "online e-books recommender system" takes an input of the user browsing history performs processing on the data collected; sorts them into sessions the keywords found in the sessions are processed further using WordNet ontology a clustering is performed on the generalised words to group them. The final recommendation stage suggests different articles to different users.

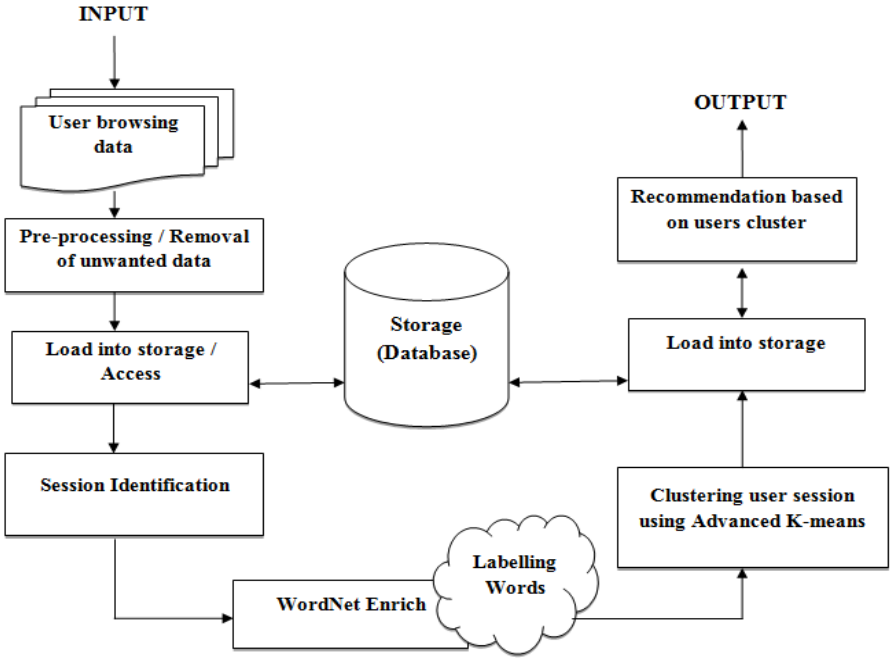

Fig 1: Proposed System

The proposed approach undergoes five steps:

1. Data Pre-processing
2. Session Identification
3. WordNet Ontology
4. Clustering User Sessions
5. Recommendation Stage

The five steps are explained in detail in the next sections.

5.1 Data Pre-processing

Various steps in the proposed system as shown above consists of firstly data pre-processing, wherein all the data about the user browsing history are collected cleaned. Cleaning such as removal of stop-words and unwanted data, then the cleaned data are stored in the database. Following, it applies several heuristics to come up with a weighting scheme that appropriately weights the keywords of each article based on information about the rest of the documents in our database. Pruning of low frequency words which may not appear in more than a small number of articles [7]. Keyword extraction, utilizing the vector space model, generates the term frequency vector, describing each article as a word-frequency pair where the words with their occurring count are measured.

5.2 Session Identification

The user browsing pattern is studied via a user's browsing history. The time when a user access a page, the duration till which the page is active tells about his likes and dislikes and in general about the behaviour of the user. Such type of information is valuable in case of recommendation. Such information can be collected from log files and clustered and at the keyword level to generate user profile clusters.

5.3 WordNet Ontology

As stated in [2] the classical text categorization system has a number of drawbacks as given below.

- The ignorance of any relation between words; thus learning algorithms are restricted to detect patterns in the used terminology only, while conceptual patterns remains ignored.
- The big dimensionality of the representation space.

A new method for text categorization is based on WordNet ontology to capture relations between words. The WordNet database is used as a text classifier; it recognises words and returns its generalised form. Hence, if two words are related comes under the same cluster with the help of WordNet.

WordNet is a thesaurus for English language based on psycholinguistics studies and developed at the University of Princeton. It was conceived as a data-processing resource which covers lexico-semantic categories called synsets. The synsets are synonyms which gather lexical items having similar significances, for example the words "a board" and "a plank" grouped in the synset {board, plank}. But "a board" can also indicate a group of people e.g., a board of directors. To disambiguate these homonymic significances a board will also belong to the synset {board, committee}. Hypernymy is a relation binding a concept to a more general concept e.g., Tulip is a flower. Here flower is a hypernym of Tulip {tulip/flower}. [13]

Table1. Relation table of the WordNet dictionary

Relation	Meaning
Synonymy	Synonyms
Antonymy	Antonyms
Hypernomy	General Concept
Hyponomy	Specific Concept
Meronymy	One of its Parts

Here in our system, we apply only the hypernomy property of the WordNet to generate more generalised form of the words. For each user session, we aggregate the keywords that make up this session. At the next step we enrich the keywords that belong to the session using related hypernyms from the WordNet database. Initially, for each given keyword of the session, we generate its graphs of hypernyms leading to the root hypernym (commonly being 'entity' for nouns).

5.4 Clustering User Sessions

Clustering user session using Advanced Kmeans, the advanced kmeans is far efficient in terms of computation time and providing the system with better clusters.

After enrichment of the words using WordNet, the words have to be clustered. The Advanced Kmeans [11] follows the following steps; this method calculates the initial centroid of the clusters in a heuristic manner.

1. For each of the data set, determine the range as the difference between the maximum and the minimum element.

2. Identify the column having the maximum range;

3. Sort the entire data set in non-decreasing order based on column having the maximum range;

4. Partition the sorted data set into 'k' equal parts.

5. Determine the arithmetic mean of each part obtained. Take these mean values as the initial centroids.

6. Repeat

Assign each data item to the cluster which has the closest centroid. Calculate new mean for each cluster.

The steps are repeated until the algorithm converges. Unlike the original kmeans in which the initial centroids are selected randomly, the algorithm determines the centroids in a meaningful way. This method consumes less time as compared to other approaches. The algorithm clusters the WordNet enriched words into groups of similar patterns.

5.5 Recommendation Stage

The recommendation engine recommends books to the user. The system goes through two phases first phase generates profile clusters and the next phase generates the recommendation to a particular user. When a user returns to the system, his cluster has already been determined, based on the recorded past sessions. It is now safe to assume that selections made by other users belonging to the same user cluster are more likely to be of interest to him/her rather than random book recommendation. Based on this simple assumption, we adjust our recommendation stage to suggest books to the user.

6. CONCLUSION

In this paper, we presented the recommender system of online books which takes as input the users' browsing data and applies pre-processing and then identifies each session and enriches the keywords found in the session with WordNet ontology and clusters the keyword using Advanced Kmeans algorithm. The users found in the same cluster are recommended same type of books.

Our proposed system is similar to the one discussed in [1] in using WordNet ontology and combining it with clustering. The system is designed for a specific purpose of recommendation of online books. Even it uses advanced kmeans algorithm to cluster as in [11]. We are combining both the approaches of clustering and recommendation for better results. Applying this model to recommend more other categories of products rather than e-books will be a part of the future work.

7. ACKNOWLEDGMENTS

I would like to express my thanks to God Almighty; my parents for their support. Sincere thanks to my project guide for reviewing my work. I am thankful to my Institution and my University.

8. REFERENCES

[1] Christos Bouras, Vassilis Tsogkas. Clustering user preferences based on W-kmeans, 2011, Seventh International Conference IEEE.

[2] Zakaria Elberrichi, Abdelattif Rahmoun, and Mohamed Amine Bentaalah. Using Wordnet for Text Categorization, the International Arab Journal of Information Technology, Vol. 5, January 2008.

[3] Feng-Hsu Wang, Hsiu-Mei Shao. Effective Personalization Recommendation Based On Time Framed Navigation Clustering and Association Mining, Effective Systems with Application 2004, (365-377)

[4] Yanjun Li, Soon M. Chung. Parallel Bisecting K-means, Wright State University USA, 2007.

[5] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom and John Riedl. GroupLens: An Open

Architecture for Collaborative Filtering of Netnews, 1994.

[6] Epimenidis Voutsakis, Giannis Varelas and Paraskevi Raftopoulou. Semantic Similarity Methods in WordNet and their Application to Information Retrieval on the Web, November 5, 2005, Bremen, Germany.

[7] Robert Cooley, Bamshad Mobasher and Jaideep Srivastava. Data preparation for mining World Wide Web, Department of Computer Science and Engineering, University of Minnesota, USA, 2000.

[8] Christos Bouras, Vassilis Poulopoulos and Vassilis Tsogkas. PeRSSonal's core functionality evaluation: Enhancing text labelling through personalized summaries, Data Knowledge and Engineering 64 (2008) 330-345.

[9] Y. Fu, K. Sandhu and M. Shih. Clustering of web users based on access patterns, Computer Science Department, University of Missouri-Rolla, 2001.

[10] Michael J. Pazzani and Daniel Billsus "Content based Recommendation Systems" Palo Alto Laboratory CA 94304.

[11] K A Abdul Naseer, S D Madhu Kumar and M P Sebastian "Enhancing the k-means algorithm by using a O(n logn) heuristic method for finding better initial centroids" 2011 Second International Conference on Emerging Applications of Information Technology, Calicut India.

[12] Lior R., Oded M."Clustering Methods" Data Mining and Knowledge Discovery Handbook, Tel Aviv University.

[13] Mahlon Lovett, "http://wordnet.princeton.edu/" Office of Communications, Princeton University, August 2014.