http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 12 (12): 2447-2453, 2013
ISSN 1812-5638 / DOI: 10.3923/itj.2013.2447.2453

A New Hybrid Schemes Combining Ontology and Clustering for Text Documents

[1]S.C. Punitha, [2]V. Thavavel and [3]M. Punithavalli
[1]Department of Computer Science and Engineering, Karunya University, Coimbatore, India
[2]Department of Computer Application, Karunya University, Coimbatore, India
[3]Department of Computer Application, Sri Ramakrishna College of Engineering, Coimbatore, India

Abstract: Data mining is a process of analyzing data from different perspectives and summarizing it into valuable information. It consist of two activities such as clustering and classification. It mainly works with numeric data, text data and the web data. Text-based algorithms have problems when dealing with different languages (synonyms, homonyms). Also, web pages contain other forms of information except text, such as images or multimedia. As a consequence, hybrid document clustering approaches have been proposed in order to combine the advantages and limit the disadvantages of the existing approaches. The main motivation behind ontology is that different people have different needs with regard to the clustering of texts. The hybrid schemes are developed using ontology and the frequent item clustering of various algorithms Ontology Based Apriori Based Clustering, Ontology based FP-Growth Based Clustering, Ontology based FP-Bonsai Clustering Algorithm have been proposed to resolve the disadvantages of existing approaches. The performance of this enhanced document clustering algorithm was tested vigorously using different datasets with performance measures to show the efficiency in clustering. Hence Ontology based FP-Bonsai Clustering Algorithm (OFPBC) shows significant improvement in terms of purity of clustering. The result shows that the datasets namely Reuters 21578,20 new Group and TDT2 which results the accuracy 0.840, 0.817 and 0.847 in OFPBC, respectively.

Key words: Document clustering, ontology, apriori algorithm, FP-growth algorithm

INTRODUCTION

Data clustering partitions is a set of unlabeled objects into combined or un-combined group of clusters. The good cluster means all the objects in a particular cluster are very similar while the objects in other clusters are different. Document clustering is a primary activity in text mining which is related to the association of documents into groups based on the area. It is very important and useful one in the information retrieval area. So while during the retrieval process, document belonging to the same cluster as the retrieved documents can also returned to the user. This could improve the recall of the information retrieval system. The useful and needed documents of the user can be retrieved using document clustering. Usually, the response of an information retrieval system is a ranked list ordered by their predictable consequence to the query. When the volume of the information database is small and the query formulated by the user is well defined, this rank list approach is efficient. But for the tremendous information source, such as the World Wide Web and the poor query condition i.e., one or two word key words, it is difficult for the retrieval system to identify the interesting item for the user that what he expected. By Applying documenting clustering to the retrieved documents, it could make easier for the users to search what they want in a short span of time. Information overload is one of the challenges in document clustering. It is estimated that more than 80% of data is stored as natural language text and finding the required information is prohibitively expensive. It results with high dimension and complex semantics. The aim of the study is to propose a model of text-clustering that can group or cluster on online news documents and a modified TF/IDF algorithm to efficiently select term/word features.

In Information retrieval systems document clustering is used to improve the precision or recall of clustering (Van Rijsbergen, 1989; Kowalski, 1997). It is also in an efficient way to find the nearest neighbors of a document (Buckley and Lewit, 1985). Based on users query the search engine results (Cutting *et al*., 1992) are browsed and coordinated by document clustering

Corresponding Author: S.C. Punitha, Department of Computer Science and Engineering, Karunya University, Coimbatore, India

(Zamir *et al*., 1997). Yahoo website provides the automatic generation of taxonomy of Web documents. (Aggarwal *et al*., 1999). The clusters in a ready existing document taxonomy (Yahoo!) is used to develop an effective document classifier for new documents and records.

Most of the available information is stored as text in the rapid growing information explosion period. Hence Data Mining (DM) and Information Retrieval (IR) from text collections (text mining) has become an active and stimulating study field. Clustering or segmentation of data is an essential data analysis procedure which has been widely premeditated across multiple disciplines for more than 40 years. The two widely used clustering techniques are generative (model-based) approaches (Cadez *et al*., 2000) and discriminative (similarity-based) approaches (Karypis *et al*., 1999). The model based clustering approaches can be used to learn generative models from the data. Each model is related to one individual cluster. Numerous algorithms available for automatic clustering of data. K Means algorithm can be applied to a set of vectors to create the clusters or groups. Usually the document is indicated by the frequency of the words which frames the record and document (the Vector space model and the Self-Organizing Semantic Map (SOM). The methods examined by Yang and Pedersen (1997) concerns the document as a container of words and it does not develop the relations between the words. The fast developing accessibility of large tracts of textual data such as blog postings, online news feeds, discussion board messages and e-mails has increased the necessity of text clustering and it becomes an active topic. On the other hand, regardless of the widespread research, unstructured clustering and textual information still remains as challenging task. Let us consider an example, the characteristics of the unstructured textual information which becomes tough for the modern clustering algorithms to describe the intrinsic structure (Gao *et al*., 2006). Data sets have unique characteristics which include more complexity to mapping upon the clustering methodology. In addition, the lack of labeled patterns in unsupervised clustering creates the partitioning task as an ill-posed crisis because there is no well known accepted methodology to develop the ideal clustering. In order to overcome these limitations, several researchers have initiated their research to examine the alternative clustering approaches.

The alternative approach incorporate background knowledge to direct each detachment task. Hence it reduce the difficulty in determining a better methodology (Hotho *et al*., 2003; Sedding and Kazakov, 2004). High dimensionality and complex semantics are the challenging problem of text clustering. Traditional clustering algorithms failed to recognize the text in a document. Whereas the hybrid schemes are developed to combine the ontology and the frequent item clustering of various algorithms to resolve the challenges of document clustering.

MATERIALS AND METHODS

General clustering using ontology: Ontology is "the specification of conceptualizations, used to help programs and humans share knowledge. Ontology is a set of concepts-such as things, events and relations that are specified in some way in order to create an agreed-upon vocabulary for exchanging information. The term "ontology" has been used for a number of years by the artificial intelligence and knowledge representation community but is now becoming part of the standard terminology of a much wider community including information systems modeling. Ontology is an explicit and formal specification of a conceptualization (Gruber, 1993). Mathematically it can be defined (Yang *et al*., 2008) as follows:

- "An ontology can be defined as an Vector O: = (C, V, P, H, ROOT), where C is the set of concepts, V (vi C) contains a set of terms and is called the vocabulary, P is the set of properties fore each concept, H is the hierarchy and ROOT is the topmost concept. Concepts are taxonomically related by the directed, acyclic, transitive, reflexive relation H C*C. H (c1, c2) shows that c1 is a subclass of c2 and for all c C it holds that H(c, ROOT)"
- Ontology defines as a common vocabulary for researchers who need to share information in a domain. It includes machine interpretable definitions of basic concepts in the domain and relations and has become common on the World-Wide Web. An example of summarization of basic ontology is shown in Fig. 1

Ontology-based document clustering: The main motivation behind ontology is that different people have different needs with regard to the clustering of texts. Empirical and mathematical analysis has shown that clustering in a high-dimensional space is very difficult and explanation why particular texts were categorized into one cluster is required. The goal of cluster analysis is the division of a set of objects into homogeneous clusters. The general steps followed by ontology-based clustering algorithms are given as:

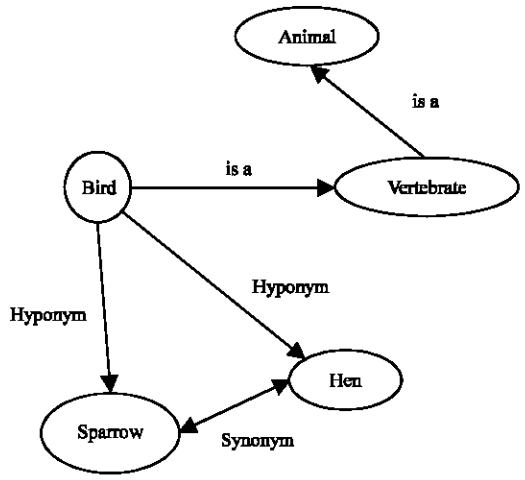

Fig. 1: Summarization of basic ontology

Step 1: Calculate distance matrix (or similarity matrix) between every pair of objects using ontology specific methods. Here, every object constitutes a separate cluster (obtaining similarity matrix)

Step 2: Using distance matrix, merge the two closest clusters (clustering process)

Step 3: Modify or rebuilt distance matrix, by treating merged clusters as one object. Methods that calculate similarity between an object and a cluster and methods that estimate similarity between clusters and ontology objects are used for this purpose (evaluation process)

Step 4: If the desired number of clusters have been reached, then stop else go to Step 2

The similarity between the objects is normally calculated using Eq. 1:

$$sim(I_i, I_j) = f_{agr}(TS(I_i, I_j), RS(I_i, I_j), AS(I_i, I_j)) \quad (1)$$

where, TS is the taxonomy similarity, RS is the relationship similarity and AS is the attribute similarity. TS are the similarity or dissimilarity between classes on the scheme and can be calculated in many ways. Some examples are measured by Wu and Palmer (1994). The idea of the relationship similarity is very simple. Similar objects should have relationships with objects that are similar to each other. When two objects O1 and O2 are compared, it should indicate all objects that have relationships with object O1 and all objects that have relationships with O2, calculate taxonomy similarity and/or attribute similarity between these two sets of objects and finally aggregate calculated similarities. The estimation of attribute similarity depends on the data types of the objects. As text documents have only strings, a lexical similarity measure is often used (Euzenat and Shvaiko, 2007). Another method is to use some distance measure like Euclidean distance or as one proposed by Manning and Schutze (1999).

Hu *et al.* (2009) found that the major problem of this ontological approach for document clustering is that it is usually difficult to find a comprehensive ontology which can cover all the concepts mentioned in a collection, especially when the documents to be clustered are from general domain. Previous study has adopted WordNet (Hotho *et al*., 2001, 2003) as the external ontology for text enrichment. However, they all have limited coverage. Another problem is that using ontology terms either as replacement or additional features has its disadvantages. While replacing original content with ontology terms may cause information loss, adding the ontology terms to the original documents vector can brings the data noise into the data set. To overcome all the disadvantages hybrid schemes are introduced to combine the ontology and frequent item clustering with various algorithm.

Ontology and frequent item clustering combine with various algorithm: Generally in English language, most of the words which have a multiple synonyms, therefore it is possible that two different documents which have no common word may represent the same topic. The frequent item based document clustering first searches the concept in document and then finds the frequent concept of Apriori algorithm (Agrawal and Srikant, 1994) Frequent Pattern followed by other algorithm like (FP)-Growth Based Clustering, FP-Bonsai Based Clustering.

Figure 2 in which uses a concept-based approach is called the ontology. Initially the model consists of identifying documents first in users query and expanding them. The preprocessing process is the process in which the unwanted documents and useless data can be removed. The preprocessing phase of the study converts the original textual data in a data-mining-ready structure, where the most significant text-features that serve to differentiate between text-categories are identified. It is the process of incorporating a new document into an information retrieval system. An effective preprocessor represents the document efficiently in terms of both space (for storing the document) and time (for processing retrieval requests) requirements and maintain good retrieval performance (precision and recall). This phase is the most critical and complex process that leads to the representation of each document by a select set of index terms. The main objective of preprocessing is to obtain the key features or key terms from online news text

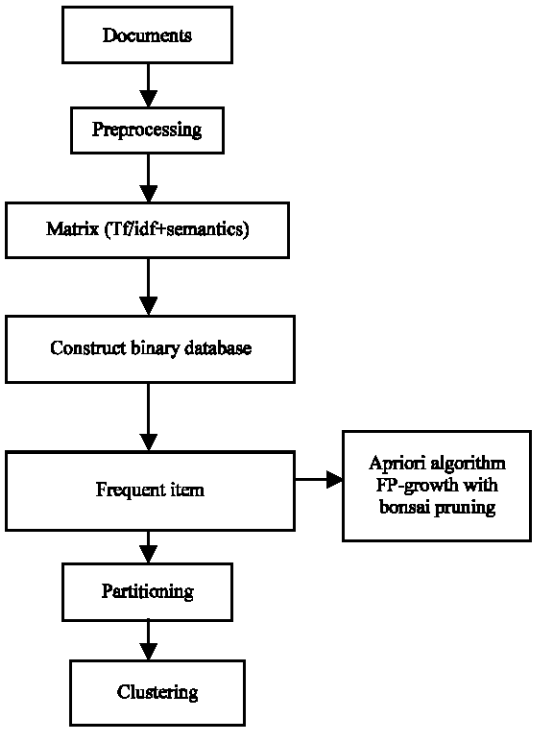

Fig. 2: Steps in frequent item based document clustering

documents and to enhance the relevancy between word and document and the relevancy between word and category.

The next thing is to determine the concept or semantic weight:

- More times the words appear in the document, more possibly it is the characteristic words
- The length of the words will also affect the importance of words. Apparently, one concept in the ontology is related to other concept in that domain ontology. That also means that the association between two concepts can be determined using the length of these two concept's connecting path (topological distance) in the concept lattice
- If the probabilities of one word is high, then the word will get additional weight
- One word may be the characteristic word even if it doesn't appear in the document
- Frequent itemset is a set of words that occur together in some minimum fraction of documents in a cluster. Then the corresponding algorithm is applied to generate a frequent itemset

Ontology based apriori based clustering (OAC): The Apriori Algorithm is the most well known association rule algorithm and it is used in most commercial products. It uses largest itemset property. "Any subset of a large itemset must be large" The basic idea of Apriori algorithm is to generate item sets of a particular size and then scan the database to count these to see if they are large. Only those candidates that are large are used to generate candidates for the next scan. Li is used to generate next Ci+1. L represents Large:

C_k: Candidate itemset of size k
I_k: Frequent itemset of size k
I_1= {Iarge 1- itemsets};
For ($k = 2$; $I_{k-1} \neq 0$; k ++) do begin
C_k = apriori-gen (I_{k1}) // New Candidates
For all transaction T∈D do begin
C_T = subset (C_k, T): // Candidates contained in T
For all candidates C ∈C_T do
C, count ++;
End
End
I_k = {c∈C_k I c, count ≥ minsub}
End

Ontology based FP-Growth Based Clustering (OFPC) and Ontology based FP-Bonsai Based Clustering (OFPBC): FP-Growth works in a divide and conquer way. It requires two scans on the database. FP-Growth first computes a list of frequent items sorted by frequency in descending order (F-List) during its first database scan. In its second scan, the database is compressed into a FP-tree. Then FP-Growth starts to mine the FP-tree for each item whose support is larger than ξ by recursively building its conditional FP-tree. The algorithm performs mining recursively on FP-tree. The problem of finding frequent itemsets is converted to searching and constructing trees recursively.

Algorithm 1 presents the pseudo code of FP-Growth (Liu *et al*., 2007):

Algorithm 1:

Procedure: FPGrowth(DB, ξ)
Define and clear F-List: F[];
foreach Transaction Ti in DB do
foreach Item a_j in T_i do
F [a_i] ++;
end
end
Sort F[];
Define and clear the root of FP-tree: *r*;
foreach *Transaction* T_i in DB do
Make T_i ordered according to F;
Call ConstructT ree($T_{i,r}$);
end
foreach item a_i in I do
Call Growth(r, a_i, ξ);
End

Later the study estimate the time complexity of computing F-List to be O using the hashing scheme (Pramudiono and Kitsuregawa, 2003). The computational cost of procedure Growth () is shown in Algorithm 2.

Algorithm 2:

Procedure: Growth(r, $a\xi$,)
if r contains a single path Z then
for each combination(denoted as γ) of the nodes in
Z do
Generate pattern $\beta = \gamma$ Ua with support = minimum support of nodes in γ;
if β. support > ξthen
Call Output (β);
end
end
else
foreach b_i in r do
Generate pattern $\beta = b_i U_a$ with support = b_i support;
if β. support > ξ then
Call Output (β);
end
Construct β's conditional database;
Construct β's conditional FP-tree Tree β;
if Tree $\beta \neq \varphi$ then
Call Growth (Treeβ, β, ξ);
end
end
end

Partitioning: Partitioning is the groups of documents which contain similar contents. For constructing initial partition (or cluster), the mined frequent itemset is used which significantly reduces the dimensionality of the text document set and clustering with reduced dimensionality is considerably more efficient and scalable. Overlapping of documents due to the use of frequent itemsets is removed as the partitions are generated directly from the frequent itemsets.

Clustering: Cluster analysis plays a vital role in many applications including document analysis. From the data mining point of view, clusters refer to similar kinds of crime in a given region of interest. Such clusters are useful in identifying similar documents. Document clustering is a more specific technique for unsupervised document organization, automatic topic extraction and fast information retrieval or filtering. Document clustering involves the use of descriptors and descriptor extraction. Descriptors are sets of words that describe the contents within the cluster. Document clustering is generally considered to be a centralized process.

EXPERIMENTAL RESULTS

The data set taken for the experimental analysis to be listed below in Table 1:

- Reuters-21578-ModApte split
- 20 News Group
- TDT2 (Topic Detection and Tracking) Dataset

The Performance Measure that can be measured by calculating the weighted average purity, Global F-measure.

Table 1: Data sets-reuters, 20 news group, TDT2

Parameters	Reuters	20 news group	TDT2
No. of documents	9603	20000	9638
No. of categories	10	20	10
Vocabulary size	22424	90833	37994
Avg. No. of words/doc	49.2	77.3	240
Avg. No. of Labels/doc	1.10	1.05	21

Table 2: Purity of cluster based on OAC, OFPC and OFPBC algorithm

Algorithm	Purity	F-measure	Time
OAC	5	5	4
OFPC	2	2	5
OFPBC	1	1	2

Table 3: F-measure of clusters for reuters 21578, 20 new groups and TDT2 dataset

Parameters	Apriori	FP growth	OAC	OFPC	OFPBC
Reuters-21578	0.781	0.810	0.802	0.833	0.840
20 new group	0.775	0.792	0.792	0.810	0.817
TDT2	0.811	0.823	0.829	0.837	0.842

The weighted average purity can be measured by the formula:

$$\rho = \sum_i \frac{ni}{n} \rho_i = \frac{n_{max}}{1} \quad (2)$$

$$\rho_i = \max_j \{\rho_{ij}\} \quad (3)$$

$$\rho_{ij} = \frac{n_{ij}}{n_i} \quad (4)$$

where, n_{ij} is the number of texts from class j in cluster I, n_i is the number of texts in cluster. n is the number of texts in the whole text set and n_{max} is the number of texts in the entire set that are part of a cluster.

To calculate Global F-measure:

$$F = \sum_i \frac{N_i}{N} \max_j (F(i,j)) \quad (5)$$

From the Table 2 the OFPBC method has a highest purity which f-measure is 1 as compared to other methods.

From the Table 3 from the experimental results, it is clear that the OFPBC (Ontology based FP-Bonsai Clustering Algorithm) shows significant improvement in terms of purity of clustering and F-Measure. For the datasets namely Reuters 21578,20 new Group and TDT2 which results 0.840, 0.817 and 0.847 in OFPBC, respectively.

CONCLUSION

Organizations and institutions around the world store data in digital form. As the number of documents grows, there is a need for robust way to extract information from them. The performance of this enhanced document

clustering algorithm was tested vigorously using different datasets shown in Table 1 and the results obtained are tabulated and discussed. From the experimental results, it is clear that the Ontology based FP-Bonsai Clustering Algorithm (OFPBC) shows significant improvement in terms of purity of clustering is shown in Table 2 and F-Measure is shown in Table 3. Hence, the OFBPC can be considered as an efficient clustering algorithm for clustering text documents. In Future Ontology based methods that combines Partition algorithm and Associative clustering can be probed. Methods that combine Apriori algorithm and FP-Bonsai algorithm can be implemented and analyzed.

REFERENCES

Aggarwal, C.C., C.S. Gates and P.S. Yu, 1999. On the merits of building categorization systems by supervised clustering. Proceedings of the 5th Conference on ACM Special Interest Group on Knowledge Discovery and Datamining, Aug. 15-18, San Diego, California, United States, pp: 352-356.

Agrawal, R. and R. Srikant, 1994. Fast algorithms for mining association rules. Proceedings of the 20th International Conference on Very Large Data Bases, September 12-15, 1994, San Francisco, USA., pp: 487-499.

Buckley, C. and A.F. Lewit, 1985. Optimization of inverted vector searches. Proceedings of the 8th Annual International SIGIR Conference on Research and Development in Information Retrieval, June 13-15, 1985, Montreal, Canada, pp: 97-110.

Cadez, I.V., S. Gaffney and P. Smyth, 2000. A general probabilistic framework for clustering individuals and objects. Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 20-23, 2000, Boston, MA, USA., pp: 140-149.

Cutting, D.R., D.R. Karger, J.O. Pedersen and J.W. Tukey, 1992. Scatter/gather: A cluster-based approach to browsing large document collections. Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, June 21-24, Copenhagen, Denmark, pp: 318-329.

Euzenat, J. and P. Shvaiko, 2007. Ontology Matching. 1st Edn., Springer-Verlag, Berlin, Germany, ISBN-13: 9783540496120, Pages: 333.

Gao, J., P.N. Tan and H. Cheng, 2006. Semi-supervised clustering with partial background information. Proceedings of the 6th SIAM International Conference on Data Mining, April 22, 2006, Bethesda, Maryland, pp: 487-491.

Gruber, T.R., 1993. A translation approach to portable ontology specifications. Knowledge Acquisit., 5: 199-220.

Hotho, A., A. Maedche and S. Staab, 2001. Text clustering based on good aggregations. Proceedings of the 2001 IEEE International Conference on Data Mining, November 29-December 2, 2001, San Jose, pp: 607-608.

Hotho, A., S. Staab and G. Stumme, 2003. Wordnet improves text document clustering. Proceedings of the SIGIR 2003 Semantic Web Workshop, July 28-August 1, 2003, Toronto, Canada, pp: 541-544.

Hu, X., X. Zhang, C. Lu and X. Zhou, 2009. Exploiting wikipedia as external knowledge for document clustering. Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, June 28-July 1, 2009, Paris, France, pp: 389-396.

Karypis, G., E.H. Han and V.K.P. Kumar, 1999. Chameleon: Hierarchical clustering using dynamic modeling. IEEE Comput., 32: 68-75.

Kowalski, G., 1997. Information Retrieval Systems: Theory and Implementation. 1st Edn., Kluwer Academic Publishers, Norwell, MA, USA., ISBN-13: 9780585320908, Pages: 282.

Liu, L., E. Li, Y. Zhang and Z. Tang, 2007. Optimization of frequent itemset mining on multiple-core processor. Proceedings of the 33rd International Conference on Very Large Data Bases, September 23-27, 2007, Vienna, Austria, pp: 1275-1285.

Manning, C. and H. Schutze, 1999. Foundations of Statistical Natural Language Processing. MIT Press, Cambridge.

Pramudiono, I. and M. Kitsuregawa, 2003. Parallel FP-growth on PC cluster. Proceedings of the 7th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, April 30-May 2, 2003, Seoul, Korea, pp: 467-473.

Sedding, J. and D. Kazakov, 2004. WordNet-based text document clustering. Proceedings of the 3rd Workshop on RObust Methods in Analysis of Natural Language Data, August 29, 2004, Geneva, Switzerland, pp: 104-113.

Van Rijsbergen, C.J., 1989. Information Retrieval. 2nd Edn., Buttersworth Publishers, London, UK., Pages: 323.

Wu, Z. and M. Palmer, 1994. Verbs semantics and lexical selection. Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics, June 27-30, 1994, Las Cruces, New Mexico, USA., pp: 133-138.

Yang, X., D. Guo, X. Cao and J. Zhou, 2008. Research on ontology-based text clustering. Proceedings of the 2008 3rd International Workshop on Semantic Media Adaptation and Personalization, December 15-16, 2008, IEEE Computer Society Washington, DC., USA., pp: 141-146.

Yang, Y.M and J. Pedersen, 1997. A comparative study on feature selection in text categorization. Proceedings of the 14th International Conference on Machine Learning, July 8-12, 1997, Nashville, TN., USA., pp: 412-420.

Zamir, O., O. Etzioni, O. Madani and R.M. Karp, 1997. Fast and intuitive clustering of web documents. Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining, Aug. 14-17, 1997, Newport Beach, California, pp: 287-290.