

College of Information Science and Technology



Drexel E-Repository and Archive (iDEA)
<http://idea.library.drexel.edu/>

Drexel University Libraries
www.library.drexel.edu

The following item is made available as a courtesy to scholars by the author(s) and Drexel University Library and may contain materials and content, including computer code and tags, artwork, text, graphics, images, and illustrations (Material) which may be protected by copyright law. Unless otherwise noted, the Material is made available for non profit and educational purposes, such as research, teaching and private study. For these limited purposes, you may reproduce (print, download or make copies) the Material without prior permission. All copies must include any copyright notice originally included with the Material. **You must seek permission from the authors or copyright owners for all uses that are not allowed by fair use and other provisions of the U.S. Copyright Law.** The responsibility for making an independent legal assessment and securing any necessary permission rests with persons desiring to reproduce or use the Material.

Please direct questions to archives@drexel.edu

A Coherent Biomedical Literature Clustering and Summarization Approach through Ontology-enriched Graphical Representations

Illhoi Yoo¹, Xiaohua Hu², and Il-Yeol Song²

¹ Department of Health Management and Informatics, School of Medicine, University of Missouri-Columbia, Columbia, MO, 65211, USA

MU.Prof.Yoo@gmail.com

² College of Information Science and Technology, Drexel University, Philadelphia, PA, 19104, USA

{thu@cis, song} @drexel.edu

Abstract. In this paper, we introduce a coherent biomedical literature clustering and summarization approach that employs a graphical representation method for text using a biomedical ontology. The key of the approach is to construct document cluster models as semantic chunks capturing the core semantic relationships in the ontology-enriched scale-free graphical representation of documents. These document cluster models are used for both document clustering and text summarization by constructing Text Semantic Interaction Network (TSIN). Our extensive experimental results indicate our approach shows 45% cluster quality improvement and 72% clustering reliability improvement, in terms of misclassification index, over Bisecting K-means as a leading document clustering approach. In addition, our approach provides concise but rich text summary in key concepts and sentences. The primary contribution of this paper is we introduce a coherent biomedical literature clustering and summarization approach that takes advantage of ontology-enriched graphical representations. Our approach significantly improves the quality of document clusters and understandability of documents through summaries.

Keywords: Document clustering, text summarization, ontology, scale-free network, MEDLINE.

1 Introduction

A huge amount of textual information has been produced and collected in text databases or digital libraries for decades because the most natural form to store information is text. For example, MEDLINE, the largest biomedical bibliographic text database, has more than 16 million articles and more than 10,000 articles are weekly added to MEDLINE. In order to tackle this pressing text information overload problem, document clustering and text summarization together have been used as a solution. This is because document clustering enables us to group similar text

information and then text summarization provides condensed text information for the similar text by extracting the most important text content from a similar document set or a document cluster. For this reason, document clustering and text summarization can be used for important components of information retrieval system. Document clustering improves information retrieval (IR) performance because similar documents grouped by document clustering tend to be relevant to the same user queries [13] [14]. Text summarization helps IR users identify which documents satisfy their needs the best by providing summaries of the retrieved documents.

In this paper, we introduce a coherent biomedical literature clustering and summarization approach. The coherence of document clustering and text summarization is required because a set of documents are usually multiple-topics. For this reason text summarization does not yield high-quality summary without document clustering. On the other hand, document clustering is not very useful for users to understand a set of documents if the explanation for document categorization or the summaries for each document cluster is not provided. In other words, document clustering and text summarization are complementary. This is the primary motivation for the coherent approach of document clustering and text summarization.

The primary contribution of this paper is we introduce a coherent biomedical literature clustering and summarization approach that takes advantage of ontology-enriched graphical representations of documents. Our approach significantly improves the quality of document clusters and understandability of documents through summaries for each document cluster.

The rest of the paper is organized as follows. Section 2 surveys the related works. In Section 3, we propose a novel graph-based document clustering approach that uses domain knowledge in an ontology and text summarization using Text Semantic Interaction Network using the semantic relationships in the document cluster model. An extensive experimental evaluation on MEDLINE articles is conducted and the results are reported in Section 4. Section 5 concludes our paper.

2 Related Works

Document Clustering: A number of document clustering approaches have been developed for several decades. Most of these document clustering approaches are based on the vector space representation and apply various clustering algorithms to the representation. Thus, the approaches can be categorized as hierarchical or partitional.

Hierarchical agglomerative clustering algorithms were used for document clustering. The algorithms successively merge the most similar objects based on the pairwise distances between objects until a termination condition holds. Thus, the algorithms can be classified by the way they select the pair of objects for calculating the similarity measure (e.g., single-link, complete-link, and average-link). An advantage of the algorithms is that they generate a document hierarchy so that users can drill up and drill down for specific topics of interest. However, due to their cubic time complexity, they are limited for a very large number of documents.

Partitional clustering algorithms (especially K-means) are the most widely-used algorithms in document clustering [10]. Most of the algorithms first randomly select k centroids and then decompose the objects into k disjoint groups through iteratively relocating objects based on the similarity between the centroids and the objects. As one of the most widely-used partitional algorithms, K-means minimizes the sum of squared distances between the objects and their corresponding cluster centroids. As a variation of K-means, BiSecting K-means [10] first selects a cluster (normally the biggest one) to split and then splits the objects into two groups (i.e. $k = 2$) using K-means. One major drawback of partitional algorithms is that clustering results are heavily sensitive to the initial centroids because the centroids are randomly selected.

Recently, Hotho et al. introduced the semantic document clustering approach that uses background knowledge [7]. The authors apply an ontology during the construction of a vector space representation by mapping terms in documents to ontology concepts and then aggregating concepts based on the concept hierarchy, which is called concept selection and aggregation (COSA). As a result of COSA, they resolve a synonym problem and introduce more general concepts in the vector space to easily identify related topics [7]. Their method, however, cannot reduce the dimensionality (i.e. the document features) in the vector space; it still suffers from the “*Curse of Dimensionality*”.

While all the approaches mentioned above represent documents as a feature vector, Suffix Tree Clustering (STC) [16] does not rely on the vector space model. STC does not treat a document as “a set of words”. One of major drawbacks of STC is that semantically similar nodes may be distant within a suffix tree, because STC does not consider the semantic relationships among phrases (nodes or base clusters). In addition, some common expressions may lead to combine unrelated documents.

Text Summarization: Text summarization has been studied since Luhn’s work in 1958 [9]. Since then, a variety of summarization approaches have been introduced. For instance, there are statistical methods based on the bag-of-words model, linguistic methods using natural language processing, knowledge-based methods using concepts and their relations, and summary generation methods. The first three approaches try to seek the most important information (usually sentences or terms) for a condensed version of documents while the last approach generates completely a new summary that consists of informative terms, phrases, clauses and sentences. The main difficulty of the last approach is to figure out how to combine them to make sentences that are grammatically correct.

In the bioinformatics/biomedical field many multi-document summarization systems have also been introduced. TextQuest [8] is designed to summarize documents retrieved in response to a keyword(s) based search on PubMed. However, it does not retain the association between the genes and the retrieved documents. MedMiner [12] can provide summarized literature information on genes but it is limited when finding relations between two genes only. In addition, it returns a few hundred sentences as the summary. Shatkey et al. [11] suggested a system, which attempts to find functional relations among genes on a genome-wide scale. However, this system requires the user to specify a representative document for each gene which describes the gene very well. Looking for the representative document may take a lot of time, effort and knowledge on the part of the user. In addition, as genes have

multiple biological functions, it is very rare to find a document that covers all aspects of a gene across various biological domains. GEISHA [3] is based on the comparison of the frequency of abstracts linked to different gene clusters. Interpretation by the end user of the biological meaning of the terms is facilitated by embedding them in the corresponding significant sentences and abstracts and by establishing relations with other, equally significant terms.

3 The Proposed Approach: CSUGAR

We present a novel coherent document clustering and summarization approach, called **Clustering and SUMmarization with GrAphical Representation for documents (CSUGAR)**. The proposed approach consists of two components, document clustering and text summarization as shown in Figure 1. Each step is discussed in detail below; see the circled numbers in Figure 1. Note the steps 1 to 3 correspond to document clustering and the steps 4 to 6 correspond to text summarization.

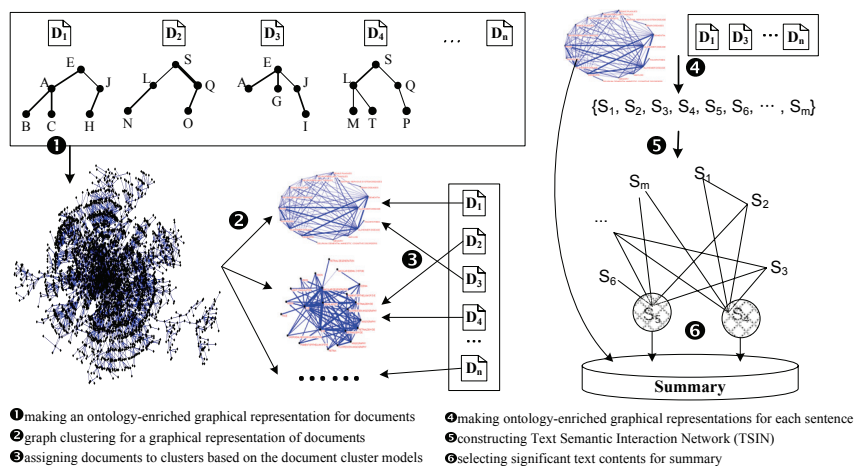


Fig. 1. The Dataflow of the CSUGAR

Step1 - Ontology-enriched Graphical Representation for Documents through Concept Mapping

The idea of the use of ontology-enriched graphical representation for documents for document clustering was first introduced in our previous work [15]. Here, we briefly introduce the graphical representation method.

The first step of all document clustering methods is to convert documents into a proper format. Since we recognize documents as a set of concepts that have their complex internal semantic relationships, we represent each document as a graph structure using the MeSH ontology. The primary motivations behind the graphical representation of documents are the following. First, the graphical representation of

documents is a very natural way to portray the contents of documents because the semantic relationship information about the concepts in documents remains on the representation while the vector space representation loses all the information. Second, the graphical representation method provides *document representation independence*. This means that the graphical representation of a document does not affect other representations. In the vector space representation, the addition of a single document usually requires the changes of every document representation. Third, the graphical representation guarantees better scalability than vector space model. Because a document representation is an actual data structure on text processing, its size should be as small as possible for better scalability. As the number of documents to be processed increases, a corpus-level graphical representation at most linearly expands or keeps its size with only some changes on edge weights, while a vector space representation (i.e. document*word matrix) at least linearly grows or increases by $n*t$ where n is the number of documents and t is the number of distinct terms in documents. For the detailed description about the graphical representation method for documents, refer to [15].

Step 2 - Graph Clustering for a Graphical Representation of Documents

A number of phenomena or systems, such as the Internet [2] have been modeled as networks or graphs. Traditionally those networks were interpreted with Erdos & Rényi's random graph theory, where nodes are randomly distributed and two nodes are connected randomly and uniformly (i.e. Gaussian distribution) [4]. However, researchers have observed that a variety of networks such as those mentioned above, deviate from the random graph theory [1] in that a few most connected nodes are connected to a high fraction of all nodes (there are a few *hub* nodes). However, these *hub* nodes cannot be explained with the traditional random graph theory. Recently, Barabasi and Albert introduced the scale-free network [2]. The scale-free network can explain the *hub* nodes with high degrees because its degree distribution decays as a power law, $P(k) \sim k^{-\gamma}$, where $P(k)$ is the probability that a vertex interacts with k other vertices and γ is the degree exponent [2].

Recently, Ferrer-Cancho and Solé have observed that the graph connecting words in English text follows a scale-free network [5]. Thus, the graphical representation of documents belongs to a highly heterogeneous family of scale-free networks. Our Scale Free Graph Clustering (SFGC) algorithm is based on the scale-free nature (i.e. the existence of a few hub vertices (concepts) in the graphical representation). SFGC starts detecting k hub vertex sets (HVSs) as the centroids of k graph clusters and then assigns the remaining vertices to graph clusters based on the relationships between the remaining objects and k hub vertex sets. For the detailed description of SFGC algorithm, refer to [15].

Step3 - Model-based Document Assignment

In this section, we explain how to assign each document to document clusters. In order to decide which document belongs to which document cluster, CSUGAR matches the graphical representation of each document with each of the graph clusters as models. Here, we might adopt graph similarity mechanisms, such as edit distance (the minimum number of primitive operations for structural modifications on a graph). However, these mechanisms are not appropriate for this task because

individual document graphs and graph clusters are too different in terms of the number of vertices and edges. As an alternative to graph similarity mechanisms we take a vote mechanism. This mechanism is based on the classification (HVS or non-HVS) of the vertices in the graph clusters according to their salient scores. This classification leads to different votes. To this end, each vertex of each individual document graph casts two different numbers of votes for document clusters based on whether the vertex belongs to HVS or non-HVS. Each document is assigned to the document cluster that has the majority of votes in the document clusters.

The next three steps correspond to text summarization. Text summarization is to condense information in a set of documents into a concise text. This text summarization problem has been addressed by selecting and ordering sentences in documents based on a salient score mechanism. We address the problem by analyzing the semantic interaction of sentences (as summary elements). This semantic structure of sentences is called Text Semantic Interaction Network (TSIN), where vertices are sentences. We select sentences (vertices in the network) as summary elements based on degree centrality. Unlike traditional approaches, we do not use linguistic features for summarization for MEDLINE abstracts since they usually consist of only single paragraphs.

Step 4 - Making Ontology-enriched Graphical Representations for Each Sentence

The first step of the graphical representation for sentences is basically the same as the graphical representation method for documents except concept extension and individual graph integration. In this step the concepts in sentences are extended using the relationships in relevant document cluster models rather than the entire concept hierarchy. In other words, we extend concepts within relevant semantic field.

Step 5 - Constructing Text Semantic Interaction Network (TSIN)

The key process of text summarization is how to select “salient” sentences (or paragraphs in some approaches) as summary elements. We assume that the sentences becoming summary have the strong semantic relationships with other sentences because summary sentences cover the main points of a set of documents and comprise a condensed version of the set. In order to represent the semantic relationship among sentences, we construct Text Semantic Interaction Network (TSIN), where vertices are sentences, edges are the semantic relationship between them, and edge weights indicate the degree of the relationships.

In order to deal with the semantic relationships between sentences and calculate the similarities (as edge weight in the network) between them, we use edit distance between the graphical representations of sentences. The edit distance between G1 and G2 is defined as the minimum number of structural modification required to become G1 into G2, where structural modification is one of vertex insertion, vertex deletion, and vertex update. For example, the edit distance between the two graphical representations of D_1 and D_2 in Figure 2 is 5.

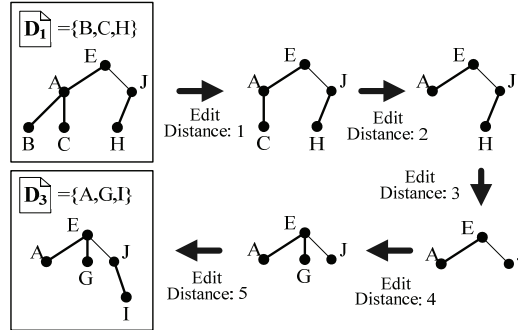


Fig. 2. Edit Distance between Two Graphical Representations of D_1 and D_2

Step 6 - Selecting Significant Text Contents for Summary

A number of approaches have been introduced to identify “important” nodes (vertices) in networks (or graphs) for decades. These approaches are normally categorized into degree centrality based approaches and between centrality based approaches. The degree centrality based approaches assume that nodes that have more relationships with others are more likely to be regarded as important in the network because they can directly relate to more other nodes. In other words, the more relationships the nodes in the network have, the more important they are. The betweenness centrality based approaches views a node as being in a favored position to the extent that the node falls on the geodesic paths between other pairs of nodes in the network [6]. In other words, the more nodes rely on a node to make connections with other nodes, the more important the node is.

These two approaches have their own advantages and disadvantages. For example, betweenness centrality based approaches yield better experiment results to find cluster centroids than other relevant approaches, while they require cubic running times so that they are not appropriate for very large graphs. Degree centrality based approaches have been criticized because they only take into account the immediate relationships for each node while they require the linear running time and provide comparable output quality with betweenness centrality based approaches.

To this end, we adopt degree centrality to measure the centrality of sentences in TSIN because of its linear computational time. In order to overcome its disadvantage, mentioned above, we measure, for each node, the semantic relationships with all other nodes (i.e., pairwise similarities for every pair of nodes) so that both immediate and distant relationships that each node has are considered while using degree centrality.

4 Experimental Evaluation

In order to measure the effectiveness of CSUGAR, we conducted extensive experiments on public MEDLINE abstracts. For the extensive experiments, first we collected document sets related to various diseases from MEDLINE. We use “MajorTopic” tag along with the disease-related MeSH terms as queries to

MEDLINE. After retrieving the base data sets, we generate various document combinations whose numbers of classes are 2 to 9 by randomly mixing the document sets. The document sets used for generating the combinations are later used as answer keys on the performance measure. For the detailed description about the document sets, the evaluation method, and the experimental setting, refer to [15].

Document Clustering

Because the full detailed experiment results are too big to be depicted in this paper, we average the clustering evaluation metric values and show the standard deviations (σ) for them to indicate how consistent a clustering approach yields document clusters (simply, the reliability of each approach). The σ would be a very important document clustering evaluation factor because document clustering is performed in the circumstance where the information about documents is unknown. Table 1 summarizes the statistical information about clustering results. From the table, we notice the following observations:

- CSUGAR outperforms the nine document clustering methods.
- CSUGAR has the most stable clustering performance regardless of test corpora, while CLUTO Bisecting K-means and K-means do not always show stable clustering performance.
- Hierarchical approaches have a serious scalability problem.
- STC and the original Bisecting K-means have a scalability problem.
- MeSH Ontology improves the clustering solutions of STC.

We observe that CSUGAR has the best performance, yields the most stable clustering results and scales very well. More specifically, CSUGAR shows 45% cluster quality improvement and 72% clustering reliability improvement, in terms of MI, over Bisecting K-means with the best parameters.

Table 1. Summary of Overall Experiment Results on MEDLINE Document Sets

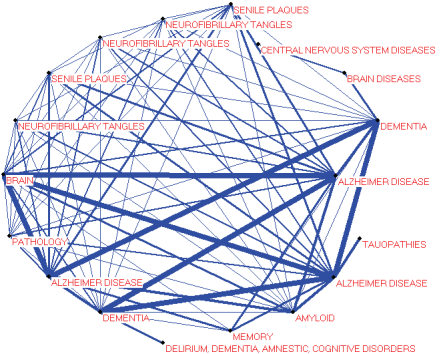
	STC		K-means	Original Bisecting K-means [10]	CLUTO Bisecting K-means		CSUGAR
	word strings	concept strings			Largest	LOS	
MI	μ : 0.429 σ : 0.238	μ : 0.359 σ : 0.149	μ : 0.128 σ : 0.148	μ : 0.395 σ : 0.193	μ : 0.161 σ : 0.139	μ : 0.096 σ : 0.112	μ : 0.053 σ : 0.031
Purity	μ : 0.601 σ : 0.214	μ : 0.731 σ : 0.098	μ : 0.932 σ : 0.080	μ : 0.666 σ : 0.154	μ : 0.918 σ : 0.064	μ : 0.944 σ : 0.056	μ : 0.947 σ : 0.030
F-measure	μ : 0.499 σ : 0.285	μ : 0.512 σ : 0.198	μ : 0.828 σ : 0.206	μ : 0.532 σ : 0.236	μ : 0.780 σ : 0.180	μ : 0.880 σ : 0.139	μ : 0.926 σ : 0.062

LOS: selecting the cluster (to be bisected) with the least overall similarity and Largest: selecting the largest cluster to be bisected. MI: the smaller, the better clustering quality. Purity and F-measure: the bigger, the better clustering quality

Text Summarization

Table 2 shows the experiment result for text summarization for a document cluster called “Alzheimer Disease”; due to the page limitation only a document cluster is presented. We believe that its document cluster model in HVS and Top 7 sentences as summary significantly help users understand the document cluster.

Table 2. Experiment Results for Text Summarization: For the Alzheimer Disease document cluster its document cluster model and key sentences as summary are shown.

<p>Document Cluster Model (HVS sets)</p>	
<p>Top 7 Sentences as Summary for the Document Cluster</p>	<ul style="list-style-type: none"> • Tau protein extracted from filaments of familial multiple system tauopathy with presenile dementia shows a minor 72-kDa band and two major bands of 64 and 68 kDa that contain mainly hyperphosphorylated four-repeat tau isoforms of 383 and 412 amino acids. • The central pathological cause of Alzheimer disease (AD) is hypothesized to be an excess of beta-amyloid (Aβ) which accumulates into toxic fibrillar deposits within extracellular areas of the brain. These deposits disrupt neural and synaptic function and ultimately lead to neuronal degeneration and dementia • In dementia of Alzheimer type (DAT), cerebral glucose metabolism is reduced in vivo, and enzymes involved in glucose breakdown are impaired in post-mortem brain tissue • Alzheimer's disease (AD), a progressive, degenerative disorder of the brain, is believed to be the most common cause of dementia amongst the elderly • The fundamental cause of Alzheimer dementia is proposed to be Alzheimer disease, i.e. the neurobiological abnormalities in Alzheimer brain • Alzheimer's disease (AD) is a degenerative disease of the brain, and the most common form of dementia • Regional quantitative analysis of NFT in brains of non-demented elderly persons: comparisons with findings in brains of late-onset Alzheimer's disease and limbic NFT dementia.

5 Conclusion

The primary contribution of this paper is we introduce a coherent biomedical literature clustering and summarization approach that takes advantage of ontology-enriched graphical representations of documents. Our approach significantly improves

the quality of document clusters and understandability of documents through summaries for each document cluster.

Acknowledgments. This research work is supported in part from the NSF Career grant (NSF IIS 0448023) NSF CCF 0514679 and the PA Dept of Health Tobacco Settlement Formula Grant (#240205, 240196).

Reference

1. Amaral, L.A.N., Scala, A., Barthélemy, M. and Stanley, H.E. *Proc. Nat. Ac. Sci USA*, 97, 2000, 11149-11152.
2. Barabasi, A.L., Albert, R. Emergence of scaling in random networks, *Science*, 286, 1999, 509.
3. Blaschke C, Oliveros JC, Valencia A (2001) Mining Functional Information Associated With Expression Arrays. *Funct. Integr. Genomics*, Vol. 1, No. 4, pp. 256-268
4. Erdos, P. and Rényi, A. On the Evolution of Random Graphs. *Publ. Math. Inst. Hungar. Acad. Sci.* 5, 1960, 17-61.
5. Ferrer-Cancho, R., and Solé, R.V., The small world of human language. In *Proceedings of the Royal Society of London*, 268, 1482, 2001, 2261–2266.
6. Hanneman, R. A., Riddle, M. 2005. Introduction to social network methods [online]. *University of California*. Available from: <http://faculty.ucr.edu/~hanneman/>
7. Hotho, A., Maedche A., and Staab S. Text Clustering Based on Good Aggregations. *Künstliche Intelligenz (KI)*, 16, 4, 2002, 48-54.
8. Iliopoulos I, Enright AJ, Ouzounis CA (2001). Textquest: document clustering of Medline abstracts for concept discovery in molecular biology. *PSB 2001*, pp. 384-395
9. Luhn, H.P. (1958) The automatic creation of literature abstracts. *IBM Journal of Research and Development*, Vol. 2, No. 2, pp. 159-165
10. Shatkey, H., Edwards, S., Wilbur, W.J. and Boguski, M. (2000) Genes, Themes and Microarrays: Using Information Retrieval For Large-Scale Gene Analysis. The 8th International Conference on Intelligent Systems Molecular Biology (ISMB 2000), La Jolla, pp. 317-328
11. Steinbach, M., Karypis, G., and Kumar, V. *A Comparison of Document Clustering Techniques*. Technical Report #00-034. University of Minnesota, 2000.
12. Tanabe, L., Scherf, U., Smith, L.H., Lee, J.K., Hunter, L., and Weinstein, J.N (1999) MedMiner: An Internet Text-Mining Tool for Biomedical Information, with Application to Gene Expression Profiling. *Biotechniques*, Vol. 27, No. 6, pp. 1210-1217
13. van Rijsbergen, C. J. *Information Retrieval*, 2nd edition, London: Butterworth, 1979.
14. Willett, P. Recent trends in hierarchical document clustering: A critical review. *Information Processing & Management*, 24, 5, 1988, 577-597.
15. Yoo I., Hu X., and Song I.Y., Clustering Ontology-enriched Graph Representation for Biomedical Documents based on Scale-Free Network Theory, accepted in the *IEEE Conference on Intelligent Systems (IEEE IS'06)*, Sept 4-6, 2006.
16. Zamir, O., and Etzioni O. Web Document Clustering: A Feasibility Demonstration, In *Proceedings of SIGIR 98*, 1998, 46-54.