The Research on Modeling Document Summarization for Audio-Based Knowledge Management

Zhigang Ji
The department of library, Hebei University of Engineering, Han dan, China

***Abstract*—There exists great text information on the internet, and there is still big unknown area in the inspection of the information. How to analyze, forecast and inspect the information will be a new issue in the field of information security. The paper, based on the calculation of semantic comparability in the how-net, proposes a series of method for calculating the semantic similarity and thus identifying the features of the text.**

***Index Terms*—B2E M-Commerce, spherical K-means clustering, structured cosine similarity (SCS), text categorization, text summarization.**

I. INTRODUCTION

Development of algorithms for automated text categorization in massive text document sets is an important research area of data mining and knowledge discovery. Most of the text-clustering methods were grounded in the term-based measurement of distance or similarity, ignoring the structure of the documents. In this paper we present a novel method named structured cosine similarity (SCS) that furnishes text clustering with a new way of modeling on document summarization, considering the structure of the documents in order to further improve the quality of clustering. This study was motivated by the problem of clustering speech documents attained from the wireless experience oral sharing conducted by mobile workforce of enterprises, fulfilling audio-based knowledge management. In other words, this problem aims to facilitate knowledge acquisition and sharing by speech.

A. B2E Mobile Commerce and Audio-Based Knowledge Management

With the advent of wireless communication technologies, the era of mobile enterprises unfold. Many international enterprises like IBM, Sun, HP, and Microsoft are vying to develop mobile enterprise servers and solution architectures. According to a Cutter report, 57% of the employees in the enterprises worldwide will be defined as "mobile workforce" by 2005 (see http://enterprise.fetnet.net/event/Special_02.htm).

Accordingly, following the e-business trend, competitive advantages built on wireless technologies in dynamic mobile environments are now widely recognized by enterprises. The quintessence of business-to-employee (B2E) mobile commerce is that enterprise employees are able to have seamless access to enterprise applications and services from any place and any time, regardless of the handset devices employed, in order to make contextualized decisions on behalf of the enterprises [1]. Among myriad kinds of enterprise applications, enterprise knowledge management is an important application in which an enterprise consciously and comprehensively gathers, organizes, shares, and analyzes its knowledge in terms of resources, documents, and people skills.

Knowledge management, however, is expensive attributable to the cost of knowledge capture, knowledge categorization, knowledge distribution, and employee's education on the creation, sharing, and use of knowledge. For instance, McKinsey & Company has long had an objective of spending 10% of its revenues on developing and managing intellectual capital [22]. In order for effective management of knowledge, hybrid solutions of people and technology are required. In other words, while compiling computerized databases of organizational knowledge, it would be highly gratifying to include "oral sharing" for subsequent knowledge categorization and distribution. Accordingly, automated speech recognition (ASR), a process of converting spoken words to speech documents for computers, plays a salient role in this aspect. The architecture of audio-based knowledge management is shown in Fig. 1. ASR is a technology that has now become one of the leading IT technologies and utilized mostly in customer service sectors of industries [2]. This technology is being currently developed and used by companies like Yahoo!, AOL, and Voice Genie. Its areas of application include voice mail processing, electronic security, and automated customer service. Speech recognition unfolds in myriad ways: discrete word recognition (recognizing only key words that the system has stored), continuous word recognition (identifying a string of numbers such as credit card numbers or zip codes), word spotting (capable of filtering out coughs and stammering and other inadvertent noises from the user), and the genre that has the ability to recognize natural speech.

doi:10.4304/jsw.5.11.1208-1215

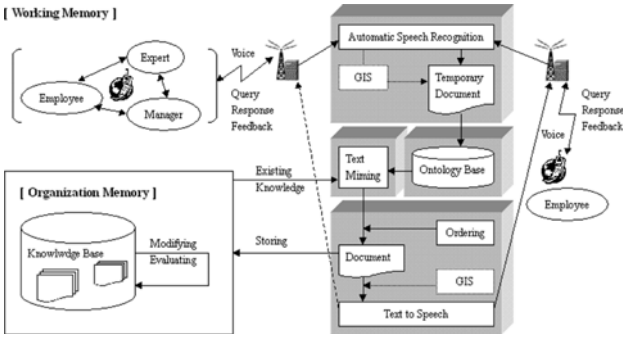


Fig. 1. Architecture of audio-based knowledge management in which audio sharing of mobile workforce is firstly processed by ASR (generating a speech document) followed by an analysis of text clustering (automatic categorization of the speech document that subsequently might be shared to the other workforce).

B. Clustering Ill-Structured Speech Documents

In spite of the advent of the ASR technology's growing maturity, speech documents attained are often short and highly ill structured/rendered, resulting in difficult knowledge categorization and distribution. (Knowledge categorization, in general, involves document summarization and summarization clustering [3].This summarization refers to intermediate-level representations of documents for facilitating subsequent clustering of documents.) Unlike web documents [that can utilize external document structures (e.g., hyperlinks [17]) or rich document features (e.g., tags, links, images, rich terms [18]) for effective document categorization] there are often neither external structures between speech documents (attained from the audio message of mobile workforce on the move) nor rich document features rendered in a speech document. Accordingly, an effective method of document summarization for speech documents is required. Excluding the consideration of external structures between documents, text categorization commonly employ vector space model [19] in which documents are summarized and represented by vectors of words (term vectors). For instance, the tfidf representation [4] summarized documents with term frequency t_{ji} inverse document. In this representation, term frequency is the number of times word i occurs in document j; the inverse document frequency g_i is low if it occurs in many documents and is highest if the word occurs in only one; a document vector is then represented as a vector of $s_{ji}=t_{ji}*g_i$.

A central problem in statistical text categorization is the high dimensionality of the feature space (one dimension for each unique word). Therefore, there is a need for the reduction of the original feature set. For instance, latent semantic indexing (LSI) [5]–[7] that further decomposed a term document matrix (resulted from the tfidf method) using a technique called singular value decomposition to construct new features as combinations of the original features, significantly reducing the high-dimensionality problem of the feature space. Moreover, LSI also considers documents that have many words in common to be semantically close, and ones with few words in common to be semantically distant.

When addressing document clustering, clustering with the complex LSI-based summarization correlates surprisingly well with how a human being, looking at documents, might classify a document collection, but suffers from heavy computation overhead [8]. Fortunately, with the simple tfidf-based summarization spherical K-means clustering (SKC) [9], [10] (evolved from the simple clustering method of K-means [11]) was proven to outperform LSI in terms of the same level of clustering quality exhibited as LSI but with less time and memory required (i.e., a significant improvement on efficiency and resource). This is because the latent concepts discovered in SKC are sparse in contrast to those in LSI being dense. Sparsity of the discovered latent concepts is important for it designating the economy or parsimony of the models constituted. Meanwhile, sparsity is crucial to computational and memory efficiency of SKC. During clustering, term vectors are clustered with certain similarity measures, such as Euclidean, cosine, extended Jaccard, etc. It is shown that Euclidean similarity is translation invariant but scale sensitive while cosine similarity is translation sensitive but scale invariant and that the extended Jaccard has aspects of both properties [12]. If clusters are to be meaningful, the similarity measure should be invariant to transformations natural to the problem domain. Furthermore, normalization may strongly affect clustering in a positive or negative way. Accordingly, a good similarity measure has to reflect the underlying semantics for the given task. With the task of text clustering, cosine similarity is a simple measure endows documents with the same composition but different sizes to be treated identically which makes this the most popular measure for clustering text documents. Moreover, due to this property, term vectors can be normalized to the unit sphere for more efficient processing as conducted in [9] and [10].

As far as we have discussed, the SKC method is comparably highly regarded due to its simplicity and good quality (in terms of fine handling of synonyms and antonyms). However, it naturally comes to a question if there is room for further improvement that can be rendered to the SKC method so as to further advance the clustering quality? This improvement is exceedingly important especially for audio-based knowledge management among mobile work force. The rationale is as follows: higher quality of performance render mobile workforce with handset devices (that are limited with resources as power, memory, screen, etc.) less overhead in sharing and accessing experience and knowledge, especially for time-critical tasks requiring decisions of quick response.

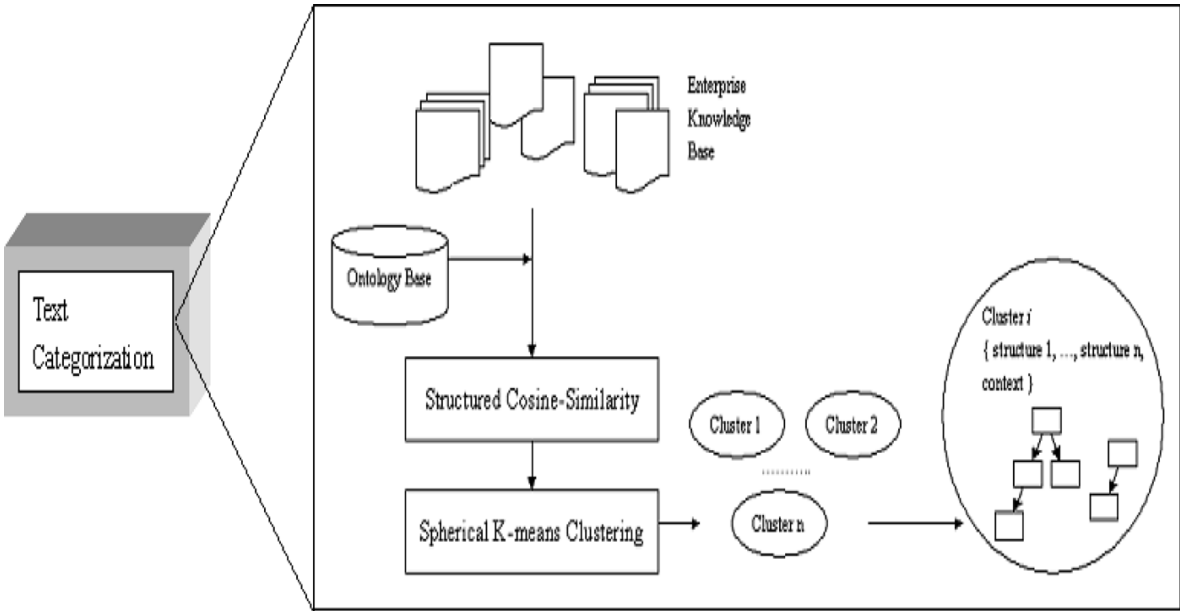


Fig. 2. Proposed method of SCS performs document summarization that is required before the application of the SKC method of summarization clustering, producing clusters containing speech documents of similar structures.

C. Research Objectives

Examining the SKC method, we discover the method disregards the whatsoever (semi) structured relations between terms. Accordingly, one way of exploring the advancement of the SKC method is to utilize these structured relations. There were works utilizing human-edited background knowledge (e.g., WordNet) to advance the performance of text clustering on synonyms or words of close semantic distance [20], [21]. (Their approaches mainly rested on query expansion with the background knowledge of synonyms.) The SKC method, on the other hand, automates building statistically synonymic relationships so as to attain text clusters each of which embodies close semantic. For the SKC method, a way to improve speech document categorization rests on a step further of clustering quality, stability, and efficiency when encountering speech documents of no rich document features.

In this paper, we present a novel methodology named structured cosine similarity, which furnishes the SKC method with a new way of modeling on document summarization by considering the structure of the documents in order to further improve the clustering quality (as shown in Fig. 2). Our evaluation results show this new modeling indeed advance the SKC method in the respects of quality, stability and efficiency. In other words, a great depth of values can be realized from ASR-generated speech documents (of neither external structures nor rich document features), shedding light on future speech-oriented applications (e.g., audio-based knowledge management), in terms of such efforts as imposing structured relations between recognized terms in documents.

II. STRUCTURED COSINE SIMILARITY

The main ideas (assumptions) underlying SCS are as follows.

• There exist certain knowledge domains in an enterprise in which the task of knowledge categorization is undertaken. A knowledge domain is presumed to be task-oriented and can be modeled by a task-oriented ontology. We currently exert as the knowledge domain troubleshooting of consumer electronics (e.g., color laser printers, digital cameras, etc.) In the application of audio-based knowledge management, mobile workforce of consumer electronic troubleshooting are able to orally share and access troubleshooting experience that is primarily underlined by certain relevant troubleshooting categorization (accordingly modeled by an electronic-troubleshooting ontology).

• In a given knowledge domain of an enterprise, there exist standard domain vocabularies that are familiar to the major classes of workforce in the enterprise.

• The ASR technique employed to generate speech documents is discrete word recognition that recognizes those key words appearing in the nodes of the task-oriented ontology. In other words, words of an oral sharing or enquiry of mobile workforce are mapped to the keywords shown in the ontology, producing a speech document. In our implementation, IBM ViaVoice is exercised to perform speech recognition and generate speech documents.

• The task-oriented ontology serves as the structure employed in modeling the new way of document summarization so as to advance the performance of the SKC method.

In the context of knowledge sharing, the term ontology means a specification of a conceptualization. That is, an ontology is a description of the concepts and relationships that can exist for an agent or a community of agents and can be designed for the purpose of enabling knowledge sharing and reuse [13]. A task-oriented ontology referred in this paper specifies concepts and causal relationships relevant to a particular task in an enterprise such as the task of consumer electronics troubleshooting (in which

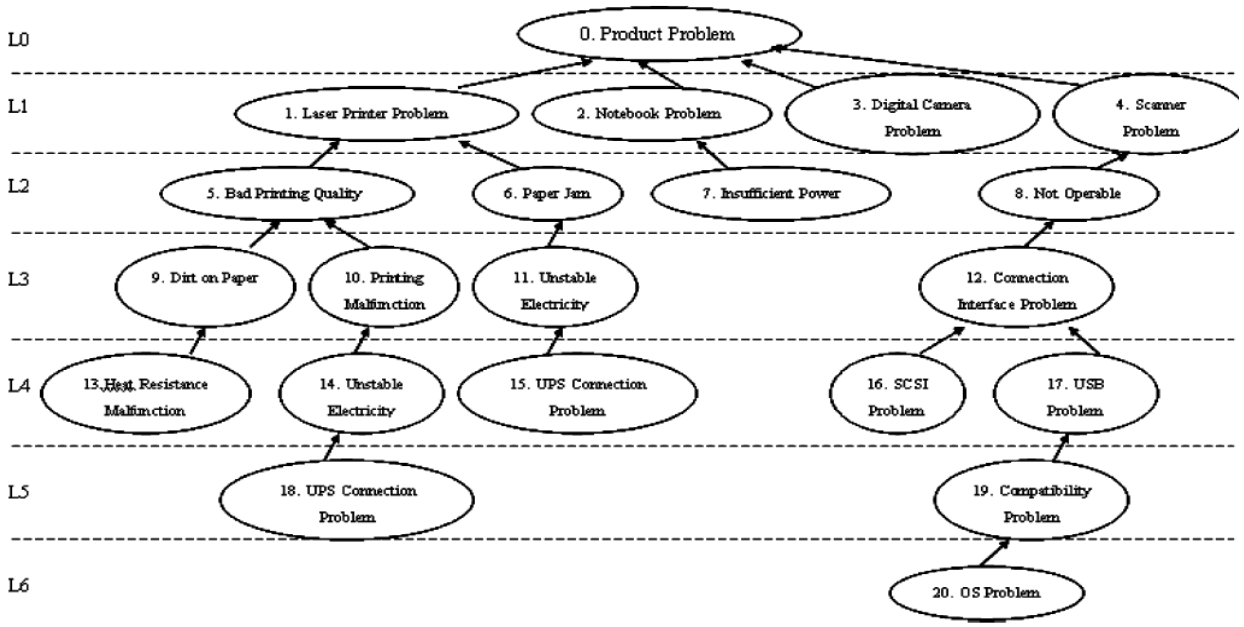


Fig. 3. Fragmented exemplar of task-oriented ontology

knowledge and experience are often related with others by troubleshooting causal relationships). Fig. 3 shows a fragmented exemplar of such task-oriented ontology, in which concepts and causal relationships are respectively represented as nodes and links that are of the type of transitive relation (i.e., if and only if, if x is related to y, and y is related to z, then x is related to z).

The following are the descriptions of nodes and links addressed in a task-oriented ontology.

• Nodes: A concept in the domain of the task. For a given concept node, its ancestor nodes of higher levels represent the causes (of the given concept) more abstract than those of lower levels. For instance, node 11 of "unstable electricity" (a cause of node 6 of "paper jam") is a cause that is more abstract than node 15 of "UPS connection problem."

• Links: Links are directional, indicating the directions of the causal relationships (in which the heads of the links represent causes while the tails denote consequences).

For instance, the cause denoted by node 15 of "UPS connection problem" often results in the consequence of node 11 of "unstable electricity." For simplicity,2 a cause for multiple consequences is represented by multiple causal links. For instance, the cause of "unstable electricity" is contributable to two consequences of "paper jam" and "printing malfunction," and thus there exist two causal links denoting this situation (a link connecting from node 11 of "unstable electricity" to node 6 of "paper jam" and another link from node 14 of "unstable electricity" to node 10 of "printing malfunction"). On the other hand, multiple causes for a single consequence (if any) are simply modeled by a causal link denoting a causal relationship between the aggregated cause and the consequence (i.e., this aggregated link either can be further refined to multiple independent links between the causes and the refined consequences or conveys that the consequence can be resulted from any cause taking from this aggregated set of causes).

The modeling of a task-oriented ontology in an enterprise can be unfolded in two phases: 1) initial construction phase (i.e., the initial development of a primitive ontology) and 2) incremental growth phase (i.e., the incremental expansion of concepts and nodes in the ontology). For the example ontology shown in Fig. 3, an initial ontology (removing the nodes of 11, 14, 15, 18) can be incrementally expanded with the addition of the nodes of "Unstable Electricity" (node 11, 14) and the nodes of "UPS Connection Problem" (node 15, 18) by the following scenario.

1) An enquiry from a mobile workforce regarding "the reason of Printing Malfunction and Paper Jam in Laser Printer" is posted to the application of audio-based knowledge management (that subsequently fails to provide any relevant solution speech documents owing to the unavailability of the relevant solutions).

2) The mobile workforce manually endeavors to discover the problems of "Unstable Electricity" and "UPS Connection Problem" accordingly.

3) This experience is orally shared by the mobile workforce to the application of audio-based knowledge management (with which the ontology-base administrator periodically recognizes the new concepts and augments the ontology in accord with the context of the domain).

III. EVALUATIONS

With the proposed method of SCS document summarization, this section accommodates further evidences that show how SCS document summarization advances the SKC method of text clustering in terms of the clustering quality. We select a troubleshooting domain of certain electronics on which its task-oriented ontology is constructed, developing a SCS demo system.

This ontology is created based on a thorough examination of the FAQ site (of 554 FAQs) of an international renowned electronics company. These FAQs are distributed as shown in Table I. The FAQs furnish us with appropriate troubleshooting knowledge in

terms of the concepts and their troubleshooting causal relationships. The resulting ontology consists of 761 concept nodes spreading over nine levels (of which a very minor portion is exemplified in Table II).

TABLE I DISTRIBUTION OF FAQS. NOTE: AN ID STANDS FOR A SYMBOL (IDENTIFIER) USED TO DENOTE THE DESIGNATED PRODUCT IN THE SUBSEQUENT EXPERIMENTS

Product	ID	Number of FAQs
Black/White Laser Printer	BL	105
DigiCam	CA	51
Color Laser Printer	CL	74
Recorder	CR	77
PDA	HA	70
Notebook	NB	43
Inkjet Printer	PR	66
Scanner	SC	68

TABLE II REPRESENTATION OF A MINOR PORTION OF THE TASK-ORIENTED ONTOLOGY EXERTED IN THE EXPERIMENTS

Item	Level	Node #	Child Node #
Product	L0	0	1、2、3、4、5、6
Digicam	L1	1	7、8、9、10、11
PDA	L1	2	12、13
Recorder	L1	3	14、15、16
Notebook	L1	4	17、18、19
Printer	L1	5	20、21、22
Scanner	L1	6	23、24、25、26、27
.....	.....	.....	.....
Software	L2	16	72、73、74、75
Hardware	L2	17	76、77、78、79、80、81、82、83

The application of the SCS demo system to the problem of clustering speech documents attained from the wireless experience oral sharing of mobile workforce then proceeds as follows: Oral sharing and enquiries of mobile workforce are first processed by ASR, attaining speech documents that are subsequently analyzed by the SCS method for knowledge categorization and distribution (so as to realize audio-based knowledge management as shown in Fig. 1).

Conduct the experiments and show the merits of SCS document summarization to the SKC method. In other words, we compare the performance of the SKC method with the three different methods of document summarization: the tfidf approach (originally employed in the SKC method), the structured imposition approach, and the s-txn approach. In order to gather comprehensive evaluation results, we conduct eight sets of experiments (in which myriad types of documents are solicited from the 554 FAQs in order to form different experiment conditions, each of which characterizes a situation of provisioned document space on which the SKC method is performed).

As follows are the eight kinds of document spaces under evaluation:

1) Documents of two categories easily differentiated but unevenly distributed (e.g., BL, NB8 or BL, CA are easily differentiated categories and thus randomly select from them the uneven amount of documents).
2) Documents of two categories easily differentiated and evenly distributed.
3) Documents of two categories not easily differentiated but evenly distributed (e.g., CL, PR or BL, CL are not easily differentiated categories).
4) Documents of two categories not easily differentiated and unevenly distributed.
5) Documents of the whole FAQs (that endow the document space of eight categories or six categories when BL, CL, PR are united together as they are all about printers).
6) Documents of the whole FAQs augmented with additional missing concepts (e.g., out of the 554 FAQs there are 423 FAQs that do not possess words appearing in the first and second level of concepts nodes of the ontology and thus can be further filled with the missing concepts in their corresponding speech documents).
7) Documents of all of the categories that are evenly distributed [e.g., the document space consists of all of the FAQs (of which categories all sizing around 65 FAQs) except documents of CA (of size 51 FAQs) and of NB (of size 43 FAQs)]. 8) Documents of all of the easily differentiated categories (e.g., by removing those documents of categories not easily differentiated such as BL and CL for PR, the document space then consists of only documents of six categories).

For the document spaces of kinds (1)–(4) documents are selected randomly according to the designated experiment settings, while for the document spaces of kinds (5)–(8) all of the 554 documents are considered and tuned in accordance with the designated experiment settings. Due to the limitation of space, we only delineate the first kind of experiments in Section III-A (Please refer to [14] for more details of the other seven kinds.) However, Section III-B will summarize the evaluation results attained from the eight kinds of experiments, accordingly manifesting the merits of SCS document summarization.

A. Experiments of the First Kind

In this set of experiments, the document space consists of documents of two categories that are easily differentiated but unevenly distributed (in terms of big difference in the amounts of their documents). Therefore, two experiments are conducted.

• Experiment (1)-1: Randomly but unevenly select documents from BL and NB, attaining 105 document of BL and 43 documents of NB. Apply the SKC clustering to generate two clusters.
• Experiment (1)-2: Randomly but unevenly select documents from BL and CA, attaining 105 documents of BL and 51 documents of CA. Apply the SKC clustering to generate two clusters.

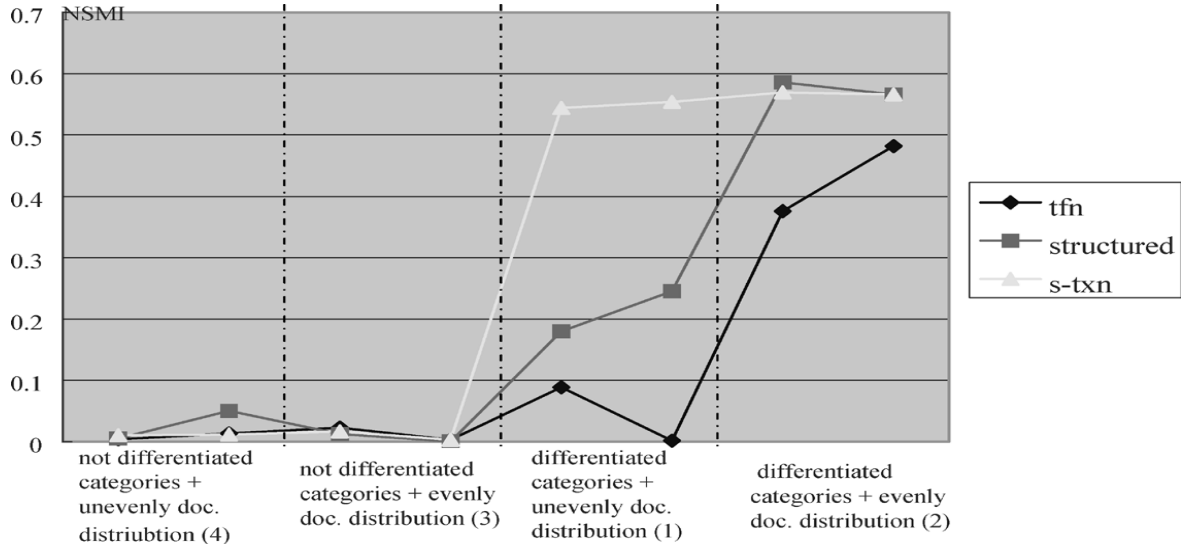


Fig. 4. Summarized results for documents of two categories. Note: "tfn" represents the SKC method equipped with its original document-summarization method of tfidf. This abbreviation will also be used in the subsequent figures

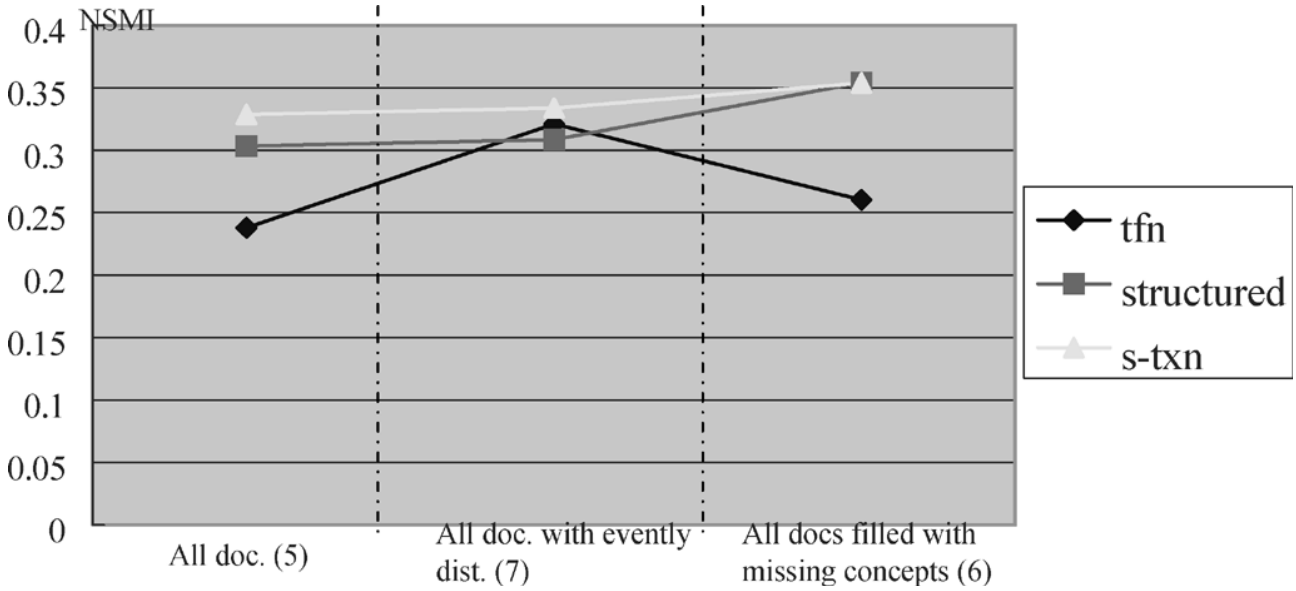


Fig. 5. Summarized results for considering all of the documents regarding document distribution

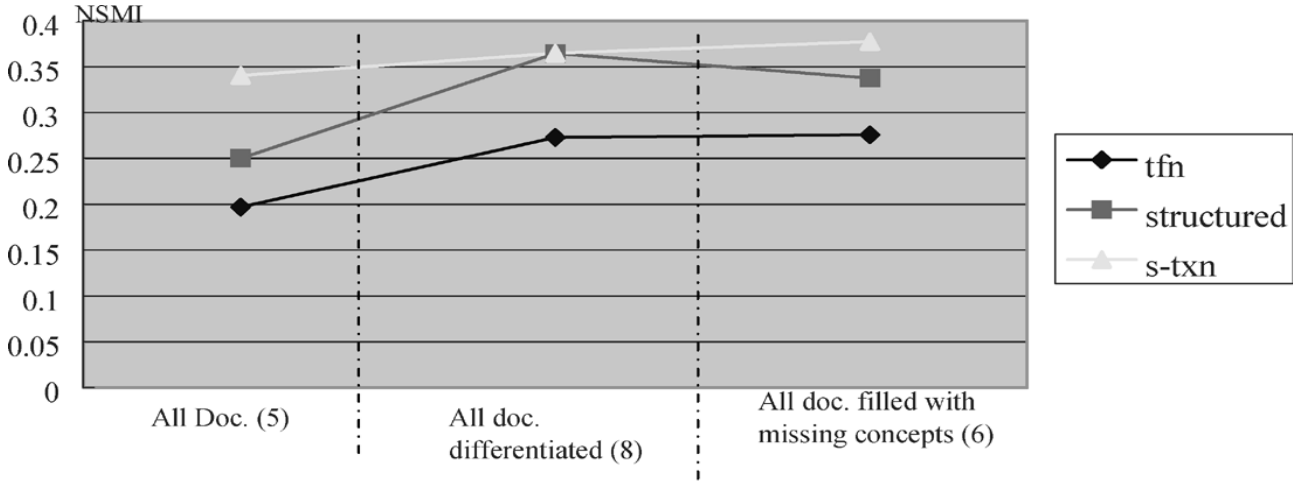


Fig. 6. Summarized results for all of the documents regarding document differentiation

B. Summary of the Results Attained From the Eight Kinds of Experiments

Since almost every kind of experiments is conducted two times (such as experiment (1)-1 and experiment (1)-2 shown in Section III-A), there are 14 sets of experiments performed. This will require a lot of space to detail. Owing to the limitation of space, this section just provides a highlight of the NSMI evaluation results indicating the overall quality of clustering (that subsequently shows the merits of the proposed method of SCS document summarization to the SKC method). These highlights are described as follows.

• From Fig. 4, with the documents of two categories not differentiated (regardless of their document amounts) we observe that the three approaches of document summarization perform around the same level of qualities (but the performance of s-txn is more stable with respect to the amounts of documents provisioned). However, the performance of the structured approach and s-txn significantly outperforms the tfidf approach when given the documents of categories differentiated. Furthermore, the performance of the tfidf approach is the least stable with respect to the amounts of the documents.

• From Fig. 5, when all of the documents are considered (and thus the number of categories considered increases), both the structured approach and s-txn outperform the tfidf approach (except in the situation of all the documents evenly distributed where the three approaches perform around the same level of qualities). Furthermore, the structured approach and s-txn benefit from the filling of the missing concepts (and thus behave semantics sound) while the tfidf approach does not.

• Fig. 6 exhibits that the structured approach behaves more like s-txn when considering all documents of categories more differentiated (in comparison with the original ones). However, s-txn exhibits a stabilized performance with respect to the increment added to the level of category differentiation.

IV. SYSTEM IMPLEMENTATION

We have implemented the system of audio-based knowledge management (exerting the SKC text categorization with the approach of s-txn document summarization) by J2EE web services (as shown in the system architecture of Fig. 7). This system has two services: Query Service and Acquisition Service [15]. Query-Service enables a mobile worker to acquire from the organization memory the most relevant knowledge to a given oral query, while Acquisition Service allows a mobile worker to render to the organization memory his/her oral working experience that would subsequently be classified into a cluster of similar documents. Upon the request of these web services, an agent residing at the mobile device of a mobile worker processes the oral inputs and manages the context of the worker in order to enable effective service communication and delivery. The PDA interfaces of Query Service and Acquisition Service are shown in Figs. 8 and 9. J2EE provides a service-oriented infrastructure to automatically support and manage distributed components called Enterprise JavaBeans (EJBs). Enterprise developers can therefore concentrate on application components. J2EE web services (EJB 2.1) include web services APIs (e.g., JAX-RPC) that can be used to communicate with other web services and allow EJBs to act as web services. When developing a web service as a stateless session bean, an endpoint interface must be defined in order to generate the JAX-RPC client stubs, a WSDL document, or both. The JAX-RPC client stub can then be packaged with the server agents in a J2EE client JAR and used to access the stateless session bean, using the SOAP protocol. If a WSDL document is generated from the endpoint interface, other SOAP toolkits such as Microsoft SOAP Type Library) can use the document to access the stateless bean. For more system demonstrations of mobile audio-based knowledge management, please see [14].

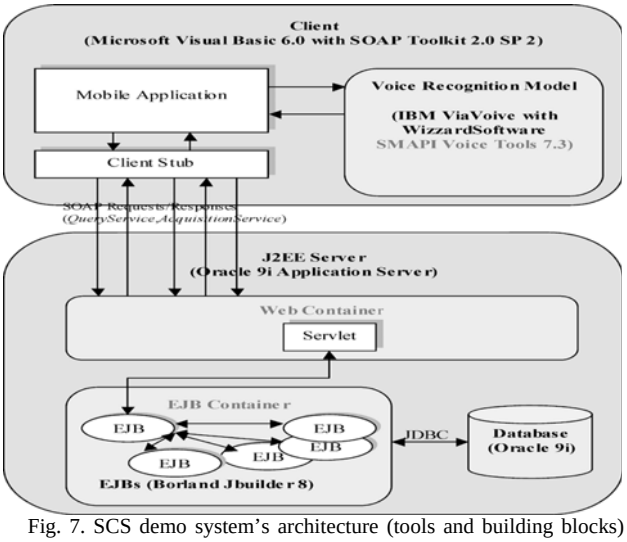


Fig. 7. SCS demo system's architecture (tools and building blocks) for audio-based knowledge management

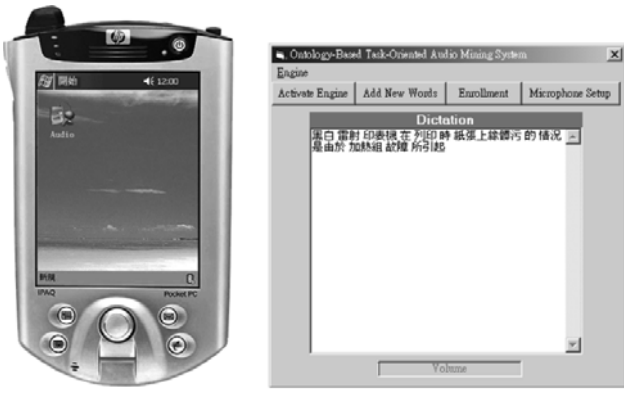


Fig. 8. Interfaces of Acquisition Service (that allows a mobile worker to render to the organization memory his/her oral working experience).

V. CONCLUSION

In this paper, we present a novel methodology named structured cosine similarity that furnishes text clustering with a new way of modeling on document summarization by considering the structure of the documents to further improve the quality of document clustering. We equip the spherical K-means clustering method (a highly regarded

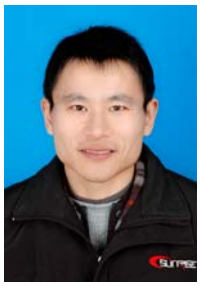
Fig. 9. Interfaces of *QueryService* (that enables a mobile worker to acquire from the organization memory the most relevant knowledge to a given oral query.

method of document clustering due to its efficiency and good quality in handling synonyms and antonyms) with SCS and attain evidences showing SCS bestows SKC a further advance in quality, stability, and efficiency.

This improvement is exceedingly important especially for the application of audio-based knowledge management among mobile work force (owing to the limited resources provisioned for the handset devices of mobile work force). Our study also sheds light on future speech-oriented applications for a great depth of values that can be realized from ASR-generated speech documents (of neither external structures nor rich document features). Besides ASR-generated text, the applicability of SCS can be extended to categorization of any texts that are short of external linking structures or rich document features but aspire for effective document categorization. Developing a suitable task-oriented ontology for a given SCS-applicable domain is better unfolded as an incremental process (as addressed in Section II-A). Nevertheless, SCS is not recommended for immense open domains owing to the enormous size of the ontology involved (e.g., WordNet). Future fruitful research includes the applicability of SCS to the other clustering methods (including the successors of the SKC method) and the extension of the application domains with our implemented web services (in addition to the application of audio-based knowledge management).

REFERENCES

[1] Ericsson Enterprise. The Path to the Mobile Enterprise. [Online]. Available: http://www.ericsson.com/products/whitepapers_pdf/whitepaper_mobile_enterprise_rc.pdf.
[2] IBM. (2001) IBMWebSphere voice server gives voice to e-business applications. [Online]. Available: http://www-3.ibm.com/software/pervasive/products/pdf/WebSphere_VS_SS_PageView.pdf.
[3] M. A. Hearst, "Text data mining: issues, techniques, and the relationship to information access," presented at the *UW/MS Workshop on Data Mining*,2008.
[4] K. Lang, "NewsWeeder: learning to Ffilter news," presented at the *Int. Conf. Machine Learning*, Tarragona, Spain,2008.
[5] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *J. Amer. Soc. Inf. Sci.*, vol. 41, no. 6, pp. 391–407,2008.
[6] S. T. Dumais, "Using LSI for information filtering: TREC-3 experiments," in *Proc. 3rd Text Retrieval Conf. (TREC3)*, Gaithersburg, MD, 1995, pp. 500–525.
[7] M. L. Littman and G. A. Keim, "Cross-language text retrieval with three languages," presented at the *Int. Conf. Neural Information Processing Systems*, Breckenridge, CO,2007.
[8] J. I. Hong. (2000) An overview of latent semantic indexing. [Online]. Available: http://www.cs.berkeley.edu/~jasonh/classes/sims240/sims-240-final-paper-lsi_files/sims-240-final-paper-lsi.doc.
[9] I. S. Dhillon, J. Fan, and Y. Guan, "Efficient clustering of very large document collections," in *Data Mining for Scientific and Engineering Applications*, R. Grossman, G. Kamath, and R. Naburu, Eds. Dordrecht, The Netherlands: Kluwer, 2007.
[10] I. S. Dhillon and D. S. Modha, "Concept decompositions for large sparse text data using clustering," *Mach. Learn.*, vol. 42, no. 1, pp. 143–175,2001.
[11] J. A. Hartigan and M. A.Wong, "AK-means clustering algorithm," *Appl. Statist.*, vol. 28, pp. 100–108, 1979.
[12] J. G. Strehl and R. Mooney, "Impact of similarity measures on web-page clustering," in *Proc. 7th Nat. Conf. Artificial Intelligence: Workshop of Artificial Intelligence for Web Search*, 2000, pp. 58–64.
[13] T. R. Gruber, "A translation approach to portable ontologies," *Knowl. Acquisit.*, vol. 5, no. 2, pp. 199–220, 1993.
[14] J. Sun and S.-T. Yuan, "Ontology-based task-oriented audio mining for Mobile B2E applications," Inf. Manage. Dept., Fu Jen Catholic Univ., Taipei, Taiwan, R.O.C., Tech. Rep., 2003.



Zhi-Gang Ji received the B.S. and M.S. degrees in information from Nanjing Agricultural University and Hebei University, China, in 1999, 2007, respectively. He was an assistant of Information Engineering at Hebei University of Engineering, Handan, China, in 1999, and became a lecturer in 2004. His research interests are in informatics, data mining, real-time systems, multimedia,and information communications.