Text Clustering using Semantic Features for Utilizing NFC Access Information

Sun Park[1], DaeKyu Kim[2] and ByungRae Cha[3]

[1]*Mokpo National University, South Korea,*
[2]*Ajantech Ltd., Seoul, South Korea,*
[3]*School of Information and Communications, GIST, South Korea*
[1]*sunpark@mokpo.ac.kr,* [2]*afoxkim@ajantech.com,* [3]*brcha@nm.gist.ac.kr*

Abstract

This paper proposes a text clustering method using the reweighted term based on semantic features for utilizing NFC content. The proposed method uses text document samples of cluster by user to reduce the semantic gap between the user's requirement and clustering results by machine for utilizing NFC access information. The method can enhance the text clustering because it uses the reweighted term which can well represent an inherent structure of text document set relevant to a user's requirement regarding NFC tags.

Keywords: *Document clustering, reweighting term, sematic feature, NMF (non-negative matrix factorization), NFC (near field communication), tags*

1. Introduction

The amount of NFC phones is increasing at a rapid pace. The NFC can be applied in many different application areas such as payment, ticketing, and information access. Generally, NFC technology can be used to replace manual typing and menu selections. It also can be used to access information with respect to the linked information of tags by user interface actions with acts of touching. The access information is composed of multimedia data. The text data of the multimedia data is played important roles for the access information [1, 2, 3]. In this article, we focus on text clustering algorithm for utilizing NFC access information.

Traditional text document clustering methods are based on bag of words (BOW) model, which represents documents with features such as weighted term frequencies (*i.e.*, vector model). However, these methods ignore semantic relationship between the terms within a document set. Recently, to overcome the problems of the vector model-based document clustering, knowledge based approaches are applied [4].

Knowledge based approaches can be either internal knowledge based or external knowledge based document clustering. Internal knowledge-based document clustering uses the inherent structure of the document set by means of a factorization technique [5-11]. These methods have been studied intensively and although they have many advantages, the successful construction of a semantic features from the original document set remains limited regarding the organization of very different documents or the composition of similar documents [1, 10]. This limitation becomes the cause of semantic gap between user's requirement and results of document clustering. External knowledge-based document clustering exploits the constructed term ontology from external knowledge database with regard to ontology as WordNet and Wikipedia [4, 12, 13, 14].

In order to enhance the internal knowledge-based approaches, this paper proposes a text document clustering method that uses the reweighted terms by semantic features of NMF and

the selected sample document of cluster by user. In addition, we propose clustering scenario regarding NFC access information using the proposed text clustering algorithm.

This paper is organized as follows: Section 2 reviews the related works regarding NFC technology, the text document clustering, and the non-negative matrix factorization (NMF); Section 3 presents the proposed text document clustering method and the proposed clustering algorithm; Section 4 shows the clustering NFC access information scenario; Section 5 shows the performance evaluation and experimental results of the proposed method. Finally, in Section 6 concludes this paper.

2. Related Works

2.1. NFC

NFC (near field communication) is a new radio technology which finds special application in the field of mobile consumer electronics for contactless communication technology. It is designed for bidirectional data transmissions over a distance of up to 10 cm and a maximum data rate of 424 kB/s. NFC technology works at an operating frequency of 13.56 MHz. NFC is standardized is ISO/IEC 18092 and ECMA-340/ECMA-352 respectively. NFC has three operating modes: peer-to-peer mode, reader/writer mode and card emulation mode. The peer-to-peer mode is an operating mode specific to NFC and allows two NFC devices to communicate directly with each other. The reader/writer mode can access contactless smartcards with regard to RFID transponders and NFC tags. The card emulation mode emulates a contactless smartcard which is can communicate with existing RFID readers [1-3].

2.2. Text Document Clustering

Internal knowledge-based document clustering uses the inherent structure of the document set by means of a factorization technique. The factorization techniques for document clustering including non-negative matrix factorization (NMF) [5-7], concept factorization (CF) [8], adaptive subspace iteration (ASI) [9], and clustering with local and global regularization (CLGR) [10] have been proposed, which can accurately identify the topics of document set from their semantic features. These methods have been studied intensively and although they have many advantages, the successful construction of a semantic features from the original document set remains limited regarding the organization of very different documents or the composition of similar documents [11]. External knowledge-based document clustering exploits the constructed term ontology from external knowledge database with regard to ontology. Recently, the term ontology techniques for document clustering are proposed such as term mutual information with conceptual knowledge by WordNet [12], concept mapping schemes from Wikipedia [4], concept weighting from domain ontology [13], and fuzzy associations with condensing cluster terms by WordNet [14], *etc*. The term ontology techniques can improve the BOW term representation of document clustering. However, it is often difficult to locate a comprehensive ontology that covers all concepts mentioned in the documents collection, which is a cause of loss of information [4, 13]. Moreover, the ontology-based method takes higher cost to construct the ontology manually by knowledge engineers and domain experts.

2.3. NMF

This section reviews NMF theory with algorithm and describes the advantage of semantic features by comparison between NMF and SVD (singular value decomposition) in Example 1. In this paper, we define the matrix notation as follows: Let X_{*j} be j'th column vector of

matrix X, X_{i*} be i'th row vector, and X_{ij} be the element of i'th row and j'th column. NMF is to decompose a given $m \times n$ matrix A into a non-negative semantic feature matrix W and a non-negative semantic variable matrix H as shown in Equation (1) [11].

$$A \approx WH \tag{1}$$

where W is a $m \times r$ non-negative matrix and H is a $r \times n$ non-negative matrix. Usually r is chosen to be smaller than m or n, so that the total sizes of W and H are smaller than that of the original matrix A.

The objective function is used minimizing the Euclidean distance between each column of A and its' approximation $\tilde{A} = WH$, which was proposed by Lee and Seung [11]. As an objective function, the Frobenius norm is used:

$$\Theta_E(W,H) \equiv \|A - WH\|_F^2 \equiv \sum_{i=1}^{m}\sum_{j=1}^{n}\left(A_{ij} - \sum_{l=1}^{r} W_{il}H_{lj}\right)^2 \tag{2}$$

Updating W and H is kept until $\Theta_E(W,H)$ converges under the predefined threshold or exceeds the number of repetition. The update rules are as follows:

$$H_{\alpha\mu} \leftarrow H_{\alpha\mu}\frac{(W^T A)_{\alpha\mu}}{(W^T WH)_{\alpha\mu}} \tag{3}$$

$$W_{i\alpha} \leftarrow W_{i\alpha}\frac{(AH^T)_{i\alpha}}{(WHH^T)_{i\alpha}} \tag{4}$$

Example 1) We illustrate an example of NMF and SVD decomposition [11, 15]. The non-negative matrix A is decomposed by *nnmf()* function of Matlab 7.8 into two non-negative matrices, W and H, as shown in Figure 1(a).

$$\underset{A}{\begin{bmatrix} 2 & 0 & 1 & 0 \\ 0 & 0 & 2 & 5 \\ 1 & 4 & 5 & 0 \\ 0 & 4 & 1 & 6 \\ 0 & 0 & 0 & 1 \end{bmatrix}} \approx \underset{W}{\begin{bmatrix} 0 & 0 & 1.692 \\ 5.022 & 0.048 & 1.573 \\ 0 & 4.128 & 4.256 \\ 5.982 & 4.076 & 0 \\ 1.001 & 0 & 0 \end{bmatrix}} \times \underset{H}{\begin{bmatrix} 0 & 0 & 0.059 & 0.998 \\ 0 & 0.980 & 0.1981 & 0.007 \\ 0.371 & 0. & 0.929 & 0 \end{bmatrix}}$$

(a) Result of NMF decomposition

$$\underset{A_{*2}}{\begin{bmatrix} 0 \\ 0 \\ 4 \\ 4 \\ 0 \end{bmatrix}} \approx 0 \times \underset{H_{12}\ W_{*1}}{\begin{bmatrix} 0 \\ 5.022 \\ 0 \\ 5.982 \\ 1.001 \end{bmatrix}} + 0.980 \times \underset{H_{22}\ W_{*2}}{\begin{bmatrix} 0 \\ 0.048 \\ 4.128 \\ 4.076 \\ 0 \end{bmatrix}} + 0 \times \underset{H_{23}\ W_{*3}}{\begin{bmatrix} 1.692 \\ 1.573 \\ 4.256 \\ 0 \\ 0 \end{bmatrix}}$$

(b) Example of column vector A_{*2} representation using semantic features and semantic variables

$$A \approx U \times S \times V$$

$$\begin{bmatrix} 2 & 0 & 1 & 0 \\ 0 & 0 & 2 & 5 \\ 1 & 4 & 5 & 0 \\ 0 & 4 & 1 & 6 \\ 0 & 0 & 0 & 1 \end{bmatrix} \approx \begin{bmatrix} 0.059 & -0.185 & 0.454 & 0.869 & -0.034 \\ 0.488 & 0.355 & 0.7138 & -0.336 & -0.117 \\ 0.454 & -0.860 & 0.023 & -0.223 & 0.067 \\ 0.739 & 0.297 & -0.529 & 0.287 & -0.067 \\ 0.079 & 0.114 & 0.063 & 0.024 & 0 \\ & & & & 988 \end{bmatrix} \times \begin{bmatrix} 9.370 & 0 & 0 & 0 \\ 0 & 5.668 & 0 & 0 \\ 0 & 0 & 2.707 & 0 \\ 0 & 0 & 0 & 1.662 \\ 0 & 0 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} 0.061 & -0.217 & 0.344 & 0.912 \\ 0.509 & -0.397 & -0.748 & 0.154 \\ 0.432 & -0.613 & 0.541 & -0.379 \\ 0.741 & 0.647 & 0.170 & 0.040 \end{bmatrix}$$

(c) Result of SVD decomposition

Figure 1. Example of NMF and SVD

Figure 1(b) shows an example of the representation of a column vector corresponding to document by a linear combination of semantic feature and semantic variable. Figure 1(c) shows the result of SVD by *svd()* function of Matlab 7.8. There are no zero values and negative values in relation to the semantic feature matrices *U* and *V* in Figure 1(c). Unlike SVD, the semantic feature matrices *W* and *H* by NMF are sparse in Figure 1(a). Intuitively, the NMF can obtain semantic features that have a small semantic range rather than SVD. In other words, the sparse property of the semantic features of NMF can cover class labels by several terms of document to be associated with the semantic features. Thus, the semantic feature can easily identify class label terms to signify document cluster. Besides, it can help to distinguish the multiple meanings of the same term [11, 15].

3. Proposed Method

This paper proposes a document clustering method using reweighting term based on semantic feature and estimation of term weighting. The proposed method consists of two phases: reweighting term and clustering document. In the subsection below, each phase is explained in full.

3.1. Reweighting Term by Semantic Features

The method of reweighting term is described as follows. First, let the number of cluster be set (it also can use to set the number of semantic feature r with connection to NMF), and then the sample documents regarding the clusters are selected by user. Second, preprocessing is performed. (*i.e.*, Rijsbergen's stop a word list is used to remove all stop words, and word stemming is removed using Porter's stemming algorithm [15, 16]. Then, the term document frequency matrixes are constructed from the selected sample documents and document set.). Finally, the reweighting term g_a^{new} is calculated by using equation (5). However, we cannot directly calculate a new weight of a'th term. In order to solve this limitation, it calculates the average weight of *a*'th row vector with regard to semantic features of document set by NMF a corresponding *a*'th term of the selected sample document .

$$g_a^{new} = g_a^{old} + \Delta g_a \tag{5}$$

Where g_a^{new} is a new weight of a'th term, g_a^{old} is a weight of *a*'th term (*i.e.*, initial value is 1.), Δg_a is variance in average weight of *a*'th row vector.

$$\Delta g_a = E(\Delta g_a^i) = \frac{1}{n}\sum_{i=1}^{n}\Delta g_a^i = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{A_{ai}}\sum_{k\in I_i}\Delta H_k W_{ak} \quad (6)$$

Where *E()* is variance, Δg_a^i is an average weight of a'th term and i'th document, *n* is the number of document in the document set, A_{ai} is a term frequency of a'th term and i'th document, I_i is term set k with respect to *i*'th variable column vector H_{*i} of document set corresponding $\Delta H_{ki} \neq 0$, ΔH_{ki} is variance in average of variable element of *k*'th term and *i*'th selected sample document.

3.2. Clustering Text Document

This section presents the clustering text document using kmeans clustering method and reweighting terms of document set. The reweighting terms are calculated by using equation (7).

$$dist(\tilde{A}_{*a}, \tilde{A}_{*b}) = 1 - csim(\tilde{A}_{*a}, \tilde{A}_{*b}) \quad (7)$$

Where $\tilde{A}$ is reweighting term document frequency matrix, *G* is weight matrix, *A* is term document frequency matrix with relation to document set.

The *k*mean algorithm takes the input parameter, *k*, and partitions a set of n objects into *k* clusters so that the resulting intra-cluster similarity is high but inter-cluster similarity is low [15, 16]. In this paper, we use cosine similarity for cluster distance measure with association to kmeans.

$$csim(\tilde{A}_{*a}, \tilde{A}_{*b}) = \frac{\sum_{i=1}^{m}\tilde{A}_{ia}\times\tilde{A}_{ib}}{\sqrt{\sum_{i=1}^{m}\tilde{A}_{ia}^2}\times\sqrt{\sum_{i=1}^{m}\tilde{A}_{ib}^2}} \quad (8)$$

Where $\tilde{A}_{*a}$ and $\tilde{A}_{*b}$ are *a*'th and *b*'th column vectors of reweighting term document frequency matrix $\tilde{A}$, respectively. These vectors have non-negative values so that are $0 \le csim() \le$ and $0 \le dist() \le 0$.

4. Clustering NFC Access Information Scenario

Figure 2 shows the clustering NFC access information scenario. Figure 2 consists of three components: the smart phone equipping with NFC, the NFC tag, the clustering server. The smart phone (*i.e.*, NFC device) has application which handles NFC access information and clustering information. The NFC tag has accessing information (*i.e.*, URL) in connection with website regarding the tag relevant information.

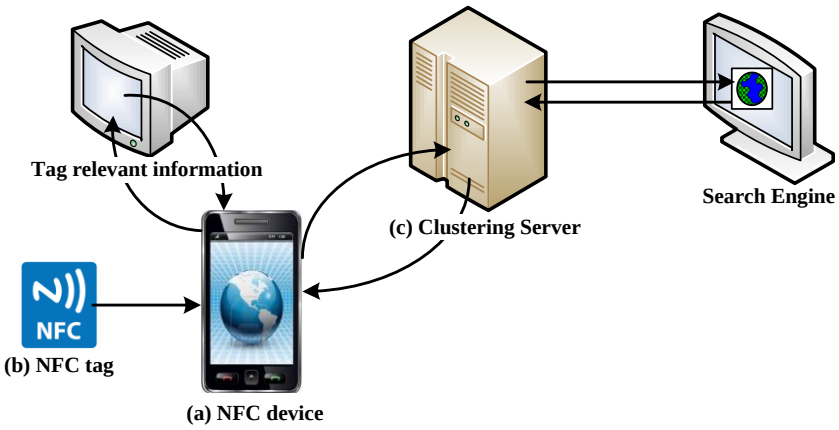

Figure 2. Scenario of Clustering NFC Access Information

In Figure 2(b), NFC device receives the tag relevant information by touching NFC tag. The tag relevant information is preprocessed for text clustering. In Figure 2(c), NFC device sends the clustering request to clustering server. The clustering server retrieves the relevant information with respect to tag content information from search engine. The clustering server clusters the retrieved information into topic groups with relation to tag information by the proposed text clustering algorithm.

5. Experiments and Evaluation

This paper uses 20 Newsgroups data set for performance evaluation [17]. To evaluate the proposed method, mixed documents were randomly chosen from the 20 Newsgroups documents. Normalized mutual information metric used to measure the document clustering performance [5-10].

Normalized mutual information metric $\overline{MI}$ as used to measure the document clustering performance [5-10]. To measure the similarity between the two sets of document clusters $C = \{ c_1, c_2, ..., c_K \}$ and $C' = \{ c'_1, c'_2, ..., c'_K \}$, the following mutual information metric MI(C,C') was used:

$$MI(C,C') = \sum_{c_i \in C, c'_j \in C'} p(c_i, c'_j) \cdot \log_2 \frac{p(c_i, c'_j)}{p(c_i) \cdot p(c'_j)} \quad (9)$$

where $p(c_i)$ and $p(c'_j)$ denote the probabilities that a document arbitrarily selected from the corpus belongs to c_i and c'_j, respectively, and $p(c_i,c'_j)$ denotes the joint probability that the selected document simultaneously belongs to c_i as well as c'_j. MI(C,C') takes values between zero and max($H(C)$,$H(C')$), where $H(C)$ and $H(C')$ are the entropies of C and C', respectively. The metric does not need to locate the corresponding counterpart in C', and the value is maintained for all permutations. The normalized metric, $\overline{MI}$, which takes values between zero and one, was used as shown in Equation (10) [5, 6, 7, 8, 9, 10, 11]:

$$\overline{MI}(C,C') = \frac{MI(C,C')}{\max(H(C), H(C'))} \quad (10)$$

The cluster numbers for the evaluation method are set by ranging from 2 to 10. For each given cluster number *K*, 50 experiments were performed on different randomly chosen clusters, and the final performance values averaged the values obtained from running experiments.

In this paper, the eight different document clustering methods are implemented. The RT, KM [15], NMF [5], CF [8], ASI [9], CLGR [10], FPCA [7], and RNMF [6] methods are document clustering methods based on internal knowledge. The KM is a document clustering using Kmeans method based on a traditional partitioning clustering technique [15]. NMF denotes Xu's method using non-negative matrix factorization [5]. CF is Xu's method using concept factorization [8]. ASI is Li's method using adaptive subspace iteration [9]. CLGR denotes Wang's method using local and global regularization [10]. FPCA is the previously proposed method using PCA (principal component analysis) and fuzzy relationship [7], and RNMF is the method proposed previously using NMF and cluster refinement [6]. The RT denotes the proposed method described within this paper.

The average normalized metric of RT is 20.8% higher than that of KM, 17.58% higher than that of NMF, 14.48% higher than that of CF, 12.88% higher than that of ASI, 7.74% higher than that of CLGR, 5.06% higher than that of FPCA, and 2.44% higher than that of RNMF.

6. Conclusion

This paper presents a text document clustering method using the reweighted term based on semantic features for enhancing document clustering. The proposed method uses document samples of cluster by user to reduce the semantic gap between the user's requirement and clustering results by machine. The method can enhance the document clustering because it uses the reweighted term which can well represent an inherent structure of document set relevant to a user's requirement. In addition, clustering scenario regarding NFC access information is proposed by using the proposed text clustering algorithm.

Acknowledgements

This work was supported by the IT R&D program of MKE/KEIT. [KI10041057 , RFID-based mobile devices to activate the service industry for mobile RFID / NFC technology convergence], This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2012R1A1A2041274).

References

[1] J. Riekki, I. Sanchez and M. Pyykkonen, "NFC-Based User Interface", in proceeding of 4th International Workshop on Near Field Communication, Helsinki, Finland, **(2012)**, pp. 3-9.
[2] R. Widmann, S. Grunberger, B. Stadlmann and J. Langer, "System Integration of NFC Ticketing into and Existing Public Transport Infrastructure", in proceeding of 4th International Workshop on Near Field Communication, Helsinki, Finland, **(2012)**, pp. 13-18.
[3] M. Roland, J. Langer and J. Scharinger, "Practical Attack Scenarios on Secure Element-enable Mobile Devices", In proceeding of 4th International Workshop on Near Field Communication, Helsinki, Finland, **(2012)**, pp. 19-24.
[4] X. Hu, X. Zhang, C. Lu, E. K. Park and X. Zhou, "Exploiting Wikipedia as External Knowledge for Document Clustering", in proceeding of KDD'09, Paris, France, **(2009)** June, pp. 389-396.
[5] W. Xu, X. Liu and Y. Gon, "Document Clustering Based On Non-negative Matrix Factorization", in proceeding of SIGIR'03, Toronto Canada, **(2003)** August, pp. 267-274.

[6] S. Park, D. U. An, B. R. Cha and C. W. Kim, "Document Clustering with Cluster Refinement and Non-negative Matrix Factorization", in proceeding of the 16th ICONIP'09, Bangkok, Thailand, **(2009)** December, pp. 281-288.
[7] S. Park and K. J. Kim, "Document Clustering using Non-negative Matrix Factorization and Fuzzy Relationship", The Journal of Korea Navigation Institute, vol. 14, no. 2, **(2010)** April, pp. 239-246.
[8] W. Xu and Y. Gong, "Document Clustering by Concept Factorization", in proceeding of the ACM SIGIR conference on research and development in information retrieval (SIGIR'04), UK, **(2004)**, pp. 202-209.
[9] T. Li, S. Ma and M. Ogihara, "Document Clustering via Adaptive Subspace Iteration", in proceeding of the ACM SIGIR conference on research and development in information retrieval (SIGIR'04), UK, **(2004)**, pp. 218-225.
[10] F. Wang and C. Zhang, "Regularized Clustering for Documents", in proceeding of the ACM SIGIR conference on research and development in information retrieval (SIGIR'07), Amsterdam, **(2007)**, pp. 95-102.
[11] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization", Nature, vol. 401, **(1999)** October, pp. 788-791.
[12] L. Jing, L. Zhou, M. K. Ng and J. Z. Huang, "Ontology-based Distance Measure for Text Clustering", in proceeding of SIAM International conference on Text Data Mining, Bethesda, MD, **(2006)**.
[13] H. H. Tar and T. T. S. Nyaunt, "Ontology-based Concept Weighting for Text Documents", World Academy of Science, Engineering and Technology, vol. 81, **(2011)**, pp. 249-253.
[14] S. Park, S. R. Lee, "Enhancing Document Clustering Using Condensing Cluster Terms and Fuzzy Association", Journal of IEICE TRANS, Information and System, vol. E94-D, no. 6, **(2011)** June, pp. 1227-1234.
[15] S. Chakrabarti, "Mining the web: Discovering Knowledge from Hypertext Data", Morgan Kaufmann Publishers, **(2003)**.
[16] W. B. Frankes and B. Y. Ricardo, "Information Retrieval, Data Structure & Algorithms", Prentice-Hall, **(1992)**.
[17] The 20 newsgroups data set. http://people.csail.mit.edu/jrennie/20Newsgroups/, **(2012)**.

Authors

Sun Park is a research professor at Institute Research of Information Science and Engineering, Mokpo National University, South Korea. He received the Ph.D degree in Computer & Information Engineering from Inha University, Korea, in 2007, the M.S. degree in Information & Communication Engineering from Hannam University, Korea, in 2001, and the B.S. degree in Computer Engineering from Jeonju University, Korea, in 1996. Prior to becoming a researcher at Mokpo National University, he has worked as a postdoctoral at Chonbuk National University, and professor in Dept. of Computer Engineering, Honam University, South Korea. His research interests include Data Mining, Information Retrieval, and Information Summarization, Convergence IT and Marine.

DaeKyu Kim received a B.S., M.S. Graduated from Chonnam University, Korea, computer engineering, in 1989, 1996. Graduated from the University Sunchon, Korea, computer engineering, in 1999, 2002. In 2012, he enrolled in a doctoral course a computer engineering Dept., Honam University, Korea. R&D Manager at company Ajantech in 2008, 2012. His main research interests include mobile communications, RFID. He is a member of kiecs.

ByungRae Cha is a research professor at school of information and communication, GIST, Korea. He received the Ph.D. degree in computer engineering from National Mokpo University in 2004 and the M.S. degree in computer engineering from Honam University in 1997. Prior to becoming a research professor at GIST, he has worked as a research professor in department of information and communication eng., Chosun University, and professor in department of computer engineering, Honam University, Korea. His research interests include Computer Security of IDS and P2P, Neural Networks Learning, Mobile-OTP, Future Internet, and Cloud Computing.