Exploring the Use of Fuzzy Signature for Text Mining

Kok Wai Wong, *Senior Member, IEEE*, Todsanai Chumwatana, *Member, IEEE*, and Domonkos Tikk

***Abstract*—The classical approaches for the traditional problems of text mining, such as document indexing, document clustering or text classification, represent the text as bag-of-words. Words, the units of the representation, are determined by tokenization, using e.g. whitespace and punctuation characters as separator. The bag-of-word based methods face problem with non-segmented text typical for some Asian languages, since the tokenization based solution cannot be applied anymore to determine the representation units. Several solutions were proposed so far, among them frequent max substring mining is adopted here because of its language-independency and favourable speed and store requirements. We present in this paper a fuzzy signature based solution using frequent max substring for non-segmented document representation, and propose how it could be applied for some typical text mining tasks. We show how the flexibility of fuzzy signatures can be exploited for text mining tasks. With the use of this proposed concept, complex decision models in text mining may be constructed more effectively in future.**

I. Introduction

With the increasing number of digital documents available in digital media and websites, it is important to find better ways to facilitate text mining. To this end, methods enabling the efficient search, retrieval, and organization of large document collections have been proposed. The three major piers of text mining tools are document indexing, document clustering and text classification algorithms. Document indexers create word-document indexes on document collection to enable efficient keyword based search. Word-document indexes assign to each word the identifiers of all documents where the word occurs. Document classification and clustering algorithms use typically also a word based representation, the so-called bag-of-words model. While at classification, the structure into which documents have to be sorted is given; at clustering no such predefined structure is provided. Classifiers assign documents into the most similar class(es), while clustering methods organize documents into groups based on their similarity. In both cases documents are compared using an appropriate similarity function based on their bag-of-word representation [1, 2]. The representation assigns weights to the words based on the relevance of words in the document and/or in the collection, oftentimes depending on the frequency of word occurrences, and the rarity of the word, etc. In essence, documents containing similar words weighted similarly will be assigned to the same class or cluster. Most of the developed techniques work well with European languages where the words, the unit of representation, can be clearly determined by simple tokenization techniques. These use whitespace and punctuation as word delimiters. Such texts are referred to as segmented text. However, some Asian languages such as Chinese and Thai are non-segmented text i.e. written continuously as a sequence of characters without explicit word boundary delimiters.

A plethora of algorithms is available also for text segmentation. Approaches can be sorted to dictionary-based, rule-based and machine learning based categories. Dictionary-based methods match each word of the dictionary against the text [3, 4] and their performance depends on the size and the quality of the dictionary. The morphology of Thai enables to use rule based techniques [5], but the accuracy then depends again on hand-crafted rules. Machine learning techniques [6] use tagged training corpora to build a statistical model able to identify boundaries between words in text. Although, this approach does not require the use of dictionary or language analysis, it still needs corpus and its performance depends critically on the characteristics of the document domain and the size of the training corpus, and also the preparation of this approach is time consuming. We remark also that for classification and clustering, the units of representation need not necessarily to coincide with meaningful dictionary units; this property is also exploited when using stemming (instead of the more complicated lemmatizers) for segmented languages.document

In [7] a frequent maximal substring (FMS) based language independent text representation was proposed that creates the representation time-efficiently and requires small storage place. The FMS-representation was first used for clustering Thai documents with SOM in [8]. The results indicate that SOM faces difficulties to separate document clusters when the clusters exhibit overlapping keywords.

Fuzzy logic could be a good alternative to help to solve the problems of overlapping in [8]. Fuzzy modelling has become very popular field in soft computing research

Manuscript received May 2, 2010. Domonkos Tikk was supported by the Alexander-von-Humboldt Foundation

K. W. Wong is with the School of Information Technology, Murdoch University, South St, WA 6150, Australia (corresponding author – phone: +618 9360 6100; fax: +61 8 9360 2941; e-mail: k.wong@murdoch.edu.au).

T. Chumwatana is with the School of Information Technology, Murdoch University, South St, WA 6150, Australia (e-mail: T.chumwatana@murdoch.edu.au)

D. Tikk is with the Institute for Computer Science, Humboldt-Univerität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany. He is on leave from Dept. of Telecommunications and Media Informatics, Budapest University of Technology and Economics, H-1117 Budapest, Magyar Tudósok krt 2., Hungary (e-mails: tikk@informatik.hu-berlin.de, tikk@tmit.bme.hu)

978-1-4244-8126-2/10/$26.00 ©2010 IEEE

because of the main feature of its ability to assign meaningful linguistic labels to the fuzzy sets in the rule base [9]. Since classical fuzzy rule based systems faced the curse of dimensionality for multivariate inputs, when the number of inputs exceeded a relatively small threshold (which naturally rise with the increasing power of computational resources), several alternatives had been proposed as remedy, such as fuzzy rule interpolation and extrapolation [10]–[13] and fuzzy signatures.

Fuzzy signature is mainly the extension of this basic concept to include fuzzy sets theory. Problems like those in the economy and medical fields normally have objects with very complex and sometimes interdependent features that need to be classified and evaluated. Fuzzy signature is thus introduced to solve complex structured data [14], [15].

One of the advantages of using fuzzy signature for complex structured decision modelling is the underlying fuzzy signature can be extracted directly from data. The constructed fuzzy signature can then be modified if necessary without changing much on the decision nature of the fuzzy signature. The objective of the paper is to explore how fuzzy signature can be used to complement the methods discussed in [7], [8]. With the nature of fuzzy signature, it could be a good alternative to solve the problem of overlapping document representation and handle uncertainties in decision making. It will be also shown that fuzzy signatures can easily incorporate additional external information about documents that could be useful for document clustering or classification.

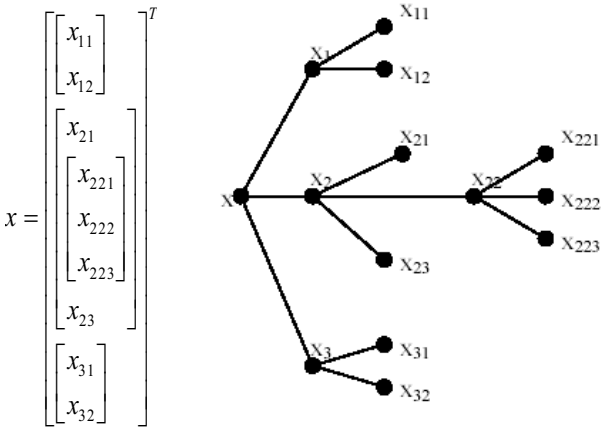


Fig. 1. Fuzzy signature. On the left the vector form of a fuzzy signature is shown, on the right the corresponding tree graph.

II. Fuzzy Signatures

Fuzzy sets were originally defined on the unit interval, $A: X \to [0,1]$, where X is the universe of discourse. As an extension, *L-fuzzy sets* were proposed by Goguen [16], $A_L: X \to L$, L being an arbitrary algebraic lattice. A practical special case of L-fuzzy sets, the *vector valued fuzzy sets* was introduced by Kóczy [17], where $A_{V,k}: X \to [0,1]^k$, and the range of membership values was the lattice of k-dimensional vectors with components in the unit interval. A further generalization of this concept is the introduction of *fuzzy signatures* and *signature sets* [14], where each vector component is possibly another nested vector as shown in Fig. 1.

Each signature corresponds to a nested vector structure or, equivalently, to a tree graph. The internal structure of the signature indicates the semantic and logical connection of state variables, corresponding to the leaves of the signature graph. The fuzzy signatures can be described as generalised vectorial fuzzy sets with possible recursive vectorial components. It can be denoted as:

$$A: X \to S^{(n)}, \tag{1}$$

where $n \geq 1$, and

$$S^{(n)} = \prod_{i=1}^{n} S_i , \tag{2}$$

$$S_i = \begin{cases} [0,1] \\ S^{(m)} \end{cases}, \tag{3}$$

and $\prod$ denotes the Cartesian product. For example a fuzzy signature A_S can be given as

$$A_S : X \to [a_i]_{i=1}^{k}, a_i = \begin{cases} [0,1] \\ [a_{ij}]_{j=1}^{k_i} \end{cases}, a_{ij} = \begin{cases} [0,1] \\ [a_{ijl}]_{l=1}^{k} \end{cases}.$$

Fuzzy signature can be considered as special multi-dimensional fuzzy data. Some of the dimensions are inter-related in the sense that they form sub-group of variables, which jointly determine some feature on a higher level.

On the example of Fig. 1, $[x_{11}\ x_{12}]$ form a sub-group that corresponds to a higher level compound variable of x_1. $[x_{221}\ x_{222}\ x_{223}]$ will then combine together to form x_{22} and $[x_{21}[x_{221}\ x_{222}\ x_{223}]x_{23}]$ is equivalent on higher level with $[x_{21}\ x_{22}\ x_{23}] = x_2$. Finally, the fuzzy signature structure will become $x = [x_1\ x_2\ x_3]$ in the example.

The relationship between higher and lower levels is governed by a set of fuzzy aggregations. The results of the parent signature at each level are computed by an appropriate aggregation of their child signatures. Let a_1 be the aggregation associating x_{11} and x_{12} used to derive x_1, thus $x_1 = x_{11}\ a_1\ x_{12}$. By referring to Figure 1, the aggregations for the whole signature structure would be a_1, a_2, a_{22}, and a_3. The aggregations a_1, a_2, a_{22}, and a_3 are not necessarily identical or different. Typical choice for aggregation can be the usual fuzzy aggregation, such as the *min* operation or any fuzzy averaging operation.

III. FREQUENT MAX SUBSTRINGS

Before the construction of the fuzzy signature that facilitates document classification and clustering on non-segmented text, the representation units of the text should be determined. In turns, we can describe non-segmented documents using vectors of these index terms. To this end, we apply frequent max substring (FMS) mining [7], which is a substring pattern mining technique applicable also on non-segmented texts where the word boundary and characteristic are not clearly defined. The goal of FMS is to provide a descriptive representation with minimal number of units, here called index terms. FMSs allow for the construction of the index structure that can be stored efficiently using trie data structure [18], [19]. FMS technique constructs only sub-trees that correspond to the frequent max substrings which contain all frequent substrings. Therefore this technique is more efficient and uses less space for storing and extracting all frequent substrings than some of the prior techniques [7]. This method uses two reduction rules: 1) a user defined frequency threshold to check extracting termination, 2) a super-substring definition to reduce the number of index terms extracted. The algorithm keeps the candidate substring in heap data structure to support computation.

Next we explain how the frequent max substrings are determined. Let we have an example string S = "GTCGTCT" and a pre-defined frequency threshold = 2

1. In the first step, the frequent substrings are determined. To this end, *n*-gram (*n*=1,...) character sequences are extracted as index terms with their frequencies and positions. Only index terms that occur at least with the pre-defined frequency will be extracted. Index terms are sorted based upon their frequency in the index data structure.
2. In the second step, the frequent max substrings are selected by keeping only those index terms that have no super-substring patterns among the index terms. This step reduces the number of the index terms significantly.

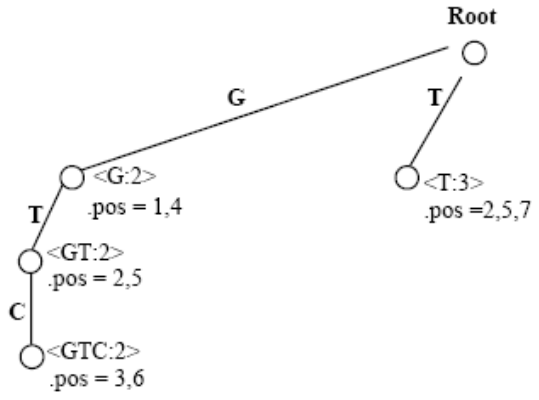


Fig. 2. The FST structure using two reduction rules

During the computation index terms are the form of a frequent suffix tree (FST) as shown in Fig. 2 for our example. Here the frequent substrings are $FSP_{S,2}$ = {<G:2>, <T:3>, <GT:2>, <GTC:2> } and the frequent max substrings, $FM_{S,2}$ = {<T:3>, <GTC:2> }. FMSs are then used as a set of indexing terms for a document.

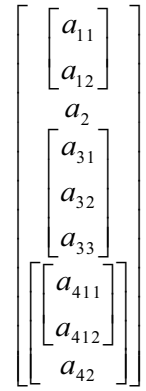
Fig. 3. Representation of the document collection with FMSs [8]

IV. DOCUMENT CLASSIFICATION USING FUZZY SIGNATURE

Let S_{S_0} denote the set of all fuzzy signatures whose structure graphs are sub-trees of the structural ("stretching") tree of a given signature S_0. Then the signature sets introduced on S_{S_0} are defined by

$$A_{S_0} : X \to S_{S_0},$$

where A_{S_0} is the membership function of the fuzzy signature S_0.



In this case, the prototype structure S_0 describes the "maximal" signature type that can be assumed by any element of X in the sense that any structural graph obtained by a set of repeated omissions of leaves from the original tree of S_0 might be the tree stretching the signature of some A_{S_0}.

There are two ways to determine the sub-trees of the fuzzy signature structure S_0. Either, it can be predetermined by a human expert of the field, or alternatively, the structure of the fuzzy signature can be determined by finding the separability from the data [20]. In this paper, the frequent max substring technique is combined with fuzzy signature in facilitating document classification and clustering. Therefore it is a combination of both, as we are incorporating human expertise and other knowledge to refine the document classification model.

In [8] a self-organized map (SOM) clustering was applied to demonstrate the applicability of FMS based representation of Thai (and in general non-segmented) texts. Document vectors were created as shown in Fig. 3, and clustered by the SOM. Clusters were then labelled for use in classification. The process is as shown in Figure 4.

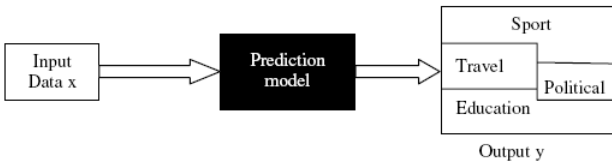


Fig. 4. Document clustering model [8]

For experimentation, in [8] a small corpus of 50 Thai documents was used as an illustration example. All documents were downloaded from Thai news websites, which consist of 15 sport documents, 15 travel documents, 15 political documents, and 5 education documents. FMSs were first generated by frequent max substring technique from the document dataset and 35 FMSs (the long and frequently occurring terms in sport, travel, political and education documents) then selected for document indexing. A sample set of the FMSs extracted is shown in Fig. 5.

1. การแข่งขัน	competition
2. นักกีฬา	athlete
3. เหรียญทอง	gold medal
4. รอบรองชนะเลิศ	Semi final round
5. ประเภทกีฬา	Sport type
6. คะแนน	Score
7. ผลการแข่งขัน	Competition result
8. ตารางการแข่งขัน	Competition timetable
12. งานไทยเที่ยวไทย	Thai travel exhibition
13. แหล่งท่องเที่ยว	Tourist attraction
14. การท่องเที่ยวแห่งประเทศไทย	The tourism authority of Thailand

Fig. 5. Sample of FMs extracted from the 50 documents

The clustering presented in [8] exhibits some deficiencies. First, it was shown SOM maps documents of different topic to the same cluster, because of the overlapping index terms occurring in the documents. For instance the groups of education and sport documents are mapped onto the same cluster because they both contain several overlapping FMSs such as **ผลการแข่งขัน** (competition result), **การจัดอันดับ** (position ranking), **ได้รับรางวัล** (getting award), etc. Handling overlapping terms via a fuzzy approach could overcome this problem. Second, no external (e.g. expert) knowledge can be incorporated easily into the model to assist the classification decision. Third, once the classification model having been established, a new model has to be built if the set of index terms is augmented with new or formerly missing FMSs.

In the following example, fuzzy signature is discussed to handle the problems presented in [8].

In this example, we have four main categories, namely Sport, Travel, Political and Education. There are two approaches for which the fuzzy signature can be constructed to recognize the four document types. First, one can construct a single fuzzy signature and use the membership function to separate the four categories. An example of the types of fuzzy signature that can be used in this case is shown in Fig. 6. However the design of the separation for the four categories could be difficult and time consuming.

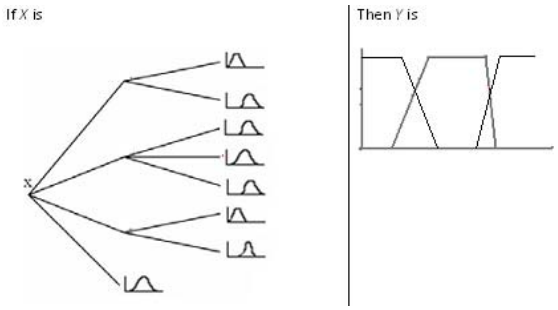


Fig. 6. Example of fuzzy signature with membership separation

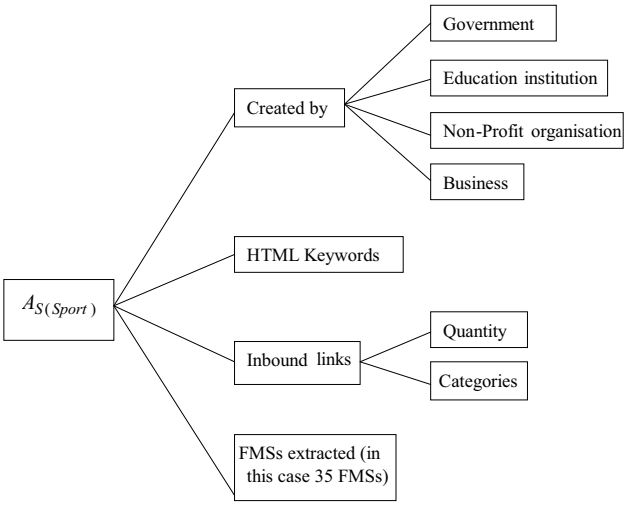


Fig. 7. Example of the Fuzzy Signature for the Sport Documents

The second method is to construct four separate fuzzy signatures, one for each type of documents namely Sport $A_{S(Sport)}$, Travel $A_{S(Travel)}$, Political $A_{S(Political)}$ and Education $A_{S(Education)}$. In each signature, the FMSs can be considered the features of the signature. If some expertise or prior knowledge is available, FMS with similar focus can be constructed as one feature with multiple leaves. The frequency of the FMSs can thus be used to construct the fuzzy sets of the leaves. With the nature of the fuzzy signature, the relationship of the different FMSs can then use any fuzzy aggregation operations to infer the results of the upper level, which eventually gives the likelihood of the document being classified for that category. From this example, it can be observed that the FMS of **การท่องเที่ยวแห่งประเทศไทย** (The tourism authority of Thailand) appears in many documents spanned across all the four categories. Under this situation, we can use the fuzzy operation like fuzzy-OR to reduce the effect of the overall decision. Alternatively, we can incorporate some extra

information to help us to make the decision more precisely. For the case of web documents, the extra information can be the document profile including authorship information, HTML-keywords, link structures, etc. An example of such fuzzy signature can be similar to the one as shown in Fig. 7.

V. CONCLUSION

In this paper, a proposal of using fuzzy signature with the frequent max substring technique has been explored. The proposal demonstrated with the benefit of the hierarchical structuring of fuzzy sets. The hierarchical structuring allows the further use of domain experts, as the information can be abstracted to higher levels analogous to patterns of human expert. By using the advantage of fuzzy signature, which is the ability to deal with problems consisting complex and interdependent features or where data is missing, an alternative document classification model can be constructed. The advantages of using fuzzy signature for knowledge management in this case are its abilities to deal with cases to handle complex structure data, to handle overlapping information, to include evolving information easily, and to handle missing information.

REFERENCES

[1] B. Liu, *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*, Berlin Heidelberg New York: Springer-Verlag, 2007.

[2] M. Steinbach, G. Karypis, and V. Kumar, "A comparison of document clustering techniques" in *Proc. of KDD Workshop on Text Mining*, 2000.

[3] V. Sornlertlamvanich, "Word Segmentation for Thai in Machine Translation System," Machine Translation, National Electronics and Computer Technology Center, Bangkok.

[4] M. R. Brent and X. Tao, "Chinese text segmentation with MBDP-1: making the most of training corpora," in *Proc. of the 39th Annual Meeting on Association for Computational Linguistics (ACL 2001)*, Toulouse, France, 2001, pp. 90–97.

[5] T. Theeramunkong, V. Sornlertlamvanich, T. Tanhermhong, and W. Chinnan, "Character-cluster based Thai information retrieval," in *Proc. of the 5th Int. Workshop on Information Retrieval with Asian Languages*, Hong Kong, 2000, pp.75–80.

[6] C. Haruechaiyasak, S. Kongyoung and C. Damrongrat, "LearnLexTo: a machine-learning based word segmentation for indexing Thai texts," in *Proc. of iNEWS'08: Proc. of the 2nd ACM workshop on Improving non English web searching ,* Napa Valley, CA, USA, 2000, pp. 85–88.

[7] T. Chumwatana, K. W. Wong and H. Xie "Thai text mining to support web search for E-commerce," in *Proc. of the 7th Int. Conf. on e-Business 2008 (INCEB2008)*, Bangkok, Thailand, 2008.

[8] T. Chumwatana, K.W. Wong, and H. Xie, "Non-segmented document clustering using self-organizing map and frequent max substring technique," in *Proc. of ICONIP 2009*, 2009, pp. 691–698.

[9] M. Sugeno, and T. Takagi, "A new approach to design of fuzzy controller," *Advances in Fuzzy Sets, Possibility Theory and Applications*, 1983, pp. 325–334.

[10] L. T. Kóczy, and K. Hirota, "Size Reduction by Interpolation in Fuzzy Rule Bases," *IEEE Transactions of System, Man and Cybernetics*, vol. 27, 1997 pp. 14–25.

[11] D. Tikk, and P. Baranyi, "Comprehensive analysis of a new fuzzy rule interpolation method," *IEEE Trans. on Fuzzy Systems*, vol 8, no. 3, 2000, pp. 281–296.

[12] S. Kovács, "New aspects of interpolative reasoning," in Proc. of the *6th Int. Conf. on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU 1996)*, Granada, Spain, 1996, pp. 477–482.

[13] Z. Huang, and Q. Shen, "Fuzzy interpolation and extrapolation: a practical approach," *IEEE Trans. on Fuzzy Systems*, vol. 16, no. 1, Feb 2008, pp. 13–28.

[14] T. Vámos, L.T. Kóczy, and G. Biró, "Fuzzy signatures in data mining," in *Proc. of the joint 9th IFSA World Congress*, 2001, pp. 2842–2846.

[15] K.W. Wong, A Chong, T.D. Gedeon, L.T. Kóczy, and T. Vámos, "Hierarchical fuzzy signatures structure for complex structured data," in *Proc. of Int. Symp. on Computational Intelligence and Intelligent Informatics 2003 (ISCIII'03)*, Nabeul, Tunisia, 2003, pp. 105–109.

[16] J.A. Goguen, "L-fuzzy Sets," *J. Math. Anal. Appl*., vol. 18, 1967, pp. 145–174.

[17] Kóczy, L. T., "Vectorial I-fuzzy Sets," in *Approximate Reasoning in Decision Analysis*, M. M. Gupta , and E. Sanchez, Eds. Amsterdam: North Holland, 1982, pp. 151–156.

[18] J. Vilo, "Discovering Frequent Patterns from Strings," Department of Computer Science. University of Helsinki, Finland, Technical Report C-1998-9, May 1998, p. 20.

[19] Paul E. Black. "*Dictionary of Algorithms and Data Structures*," NIST Available: http://www.nist.gov/dads/HTML/trie.html

[20] K.W. Wong, T.D. Gedeon, and L.T. Kóczy, "Construction of fuzzy signature from data: an example of SARs pre-clinical diagnosis system," in *Proc. of IEEE Int. Conf. on Fuzzy Systems (FUZZ-IEEE 2004)*, Budapest, Hungary, 2004, paper 1353.