Towards An Intuitionistic Agglomerative Fuzzy Hierarchical Clustering Algorithm for Music Recommendation in Folksonomy

Chun Guan
Department of Computer Science
University of Liverpool
Liverpool, UK
chunguan@liverpool.ac.uk

Kevin Kam Fung Yuen
Department of Computer Science and Software Engineering
Xi'an Jiaotong-Liverpool University
Suzhou, China
kevinkf.yuen@gmail.com
kevin.yuen@xjtlu.edu.cn

Frans Coenen
Department of Computer Science
University of Liverpool
Liverpool, UK
coenen@liverpool.ac.uk

***Abstract*—Folksonomy, a system for social tagging or collaborative tagging, is popular in Semantic Web Research. Folksonomy is applied to items, such as music pieces, which their personalized tags can be annotated by users. Recommendation systems can use these tags to produce meaningful information. Clustering methods, such as the agglomerative hierarchical clustering (AHC) method, can be applied in the context of recommendation system. This paper proposes the Intuitionistic Fuzzy Agglomerative Hierarchical Clustering (IFAHC) algorithm for recommendation using social tagging. The Intuitionistic Fuzzy Set (IFS) concept is used to represent tag values which are vague and uncertain. IFAHC can cluster items represented by using IFS into different groups. The application of IFAHC to music recommendation is used to demonstrate the usability of the proposed method.**

Keywords- Agglomerative Hierarchical Clustering; Intuitionistic Fuzzy Set; Social Tagging System; Folksonomy; Music Recommendation.

I. INTRODUCTION

Clustering analysis [1, 2] is an important technique in data mining. Hierarchical clustering [1-6] is a classical and popular clustering algorithm since it was proposed in 1963 [1]. Agglomerative Hierarchical Clustering (AHC) algorithms have been progressively applied in many areas, as described in [3-6]. AHC is the method to build a bottom-to-top hierarchical decomposition of the data set on the basis of dissimilarities between objects. The clustering result of an AHC algorithm is typically illustrated using a dendrogram offering easy interpretation by a decision makers.

Social tagging and collaborative tagging systems were first described using the term "folksonomy" by Vander [7]. Folksonomy allows users to annotate their favorite resources and items with personalized tags [6]. Clustering algorithms have been applied to organize the large collections of data using folksonomy [6, 8, 9]. Resources and items can be divided into clusters according to their tags. The clustering results can be used in a recommendation system. Since the tags are personalized, defined by users with freely chosen vocabularies, the tag values tend to contain fuzziness [10].

To deal with this fuzziness, the Intuitionistic Fuzzy Set (IFC) concept described in [11, 14] is used to represent the tag values. Fuzzy set theory, used to represent the uncertainty membership of items to groups, was established some fifty years ago [12, 13]. By adding the non-membership concept the idea of Intuitionistic Fuzzy Sets (IFS) was developed as the extension of fuzzy set. IFS has been applied in various areas such as decision making, machine learning and pattern recognition.

This paper proposes to combine IFS and AHC to form Intuitionistic Fuzzy Agglomerative Hierarchical Clustering (IFAHC) to cluster items according to their IFS values so as to give meaningful patterns. In this paper the proposed IFAHC is used to cluster music pieces according to a folksonomy tagging process. In IFAHC, IFS is applied to represent items with both their membership degrees and their non-membership degrees.

The rest of this paper is organized as follows. Section II describes the proposed IFAHC algorithm. Section III presents a demonstration of applying IFAHC to music recommendation. Section IV gives a summary of this research and further research directions.

II. INTUITIONISTIC FUZZY AGGLOMERATIVE HIERARCHICAL CLUSTERING

A. Object representation in IFS

The definition of IFS is given as follows with respect to [11, 14]. Let E be a fixed set. An intuitionistic fuzzy set A in X is expressed as below.

$$A = \{\langle x, \mu_A(x), \nu_A(x) \big| x \in E \big| \rangle\} \quad (1)$$

where $\mu_A(x)$ is the membership degree of x, i.e. $\mu_A : X \to [0,1]$, $\nu_A(x)$ is the non-membership degree of x, i.e. $\nu_A : X \to [0,1]$. $\mu_A(x)$ and $\nu_A(x)$ satisfy the condition below.

$$0 \le \mu_A(x) + \nu_A(x) \le 1 \qquad \forall x \in X \quad (2)$$

The intuitionistic index $\pi_A(x)$ represents the hesitancy degree of x to A as below.

$$\pi_A(x) = 1 - \mu_A(x) - \nu_A(x) \quad (3)$$

Objects in a folksonomy are labeled by user defined tags, and the values of some tags may be fuzzy. This paper proposes an approach to representing object tag values using IFS.

Assume that an object x has been tagged by M users using a collection of tags, T. An expert will pick up a tag t from collection T as an attribute of x for clustering. Three kinds of relationships between object x and tag t can exist.

- The relationship that a user tagged x by t, which is similar to the membership relationship, could be represented by $\mu_A(x)$.
- The relationship that a user did not tag x by t, which is similar to the non-membership relationship, could be represented by $\nu_A(x)$.
- The relationship that a user tagged x by t' ($t' \neq t$, but t is similar to t', and $t' \in T$), which is similar to the hesitancy relationship, could be represented by $\pi_A(x)$. The similarity can be defined by the overlap of keywords.

Let M_μ, M_ν and M_π be the numbers of users with respect to the above three relationships respectively, and $M_\mu + M_\nu + M_\pi = M$. The equations below are defined for transforming the tagged objects into IFSs.

$$\mu_t(x) = \frac{M_\mu}{M} \quad (4)$$

$$\nu_t(x) = \frac{M_\nu}{M} \quad (5)$$

$$\pi_t(x) = \frac{M_\pi}{M} \quad (6)$$

Finally, the objects can be transformed to a collection of IFSs of the above form.

B. Objects clustering

Agglomerative hierarchical clustering [3] is a bottom-up strategy. [15] briefly described the hierarchical clustering methods in terms of three steps. The method starts by initializing each object as an atomic cluster and then merges them into larger clusters, until all objects are in a single cluster or termination condition is satisfied [2]. The proposed IFAHC operates in a somewhat different manner to classical AHC. The operation is as follows.

1) Initialize each object as an individual cluster.

2) Determine dissimilarities between clusters.

Euclidean distance has been used to calculate the dissimilarities in classical AHC. In IFAHC, normalized Euclidean distance for IFS, as proposed by [14], was used to calculate the dissimilarities:

$$d(A,B) = \sqrt{\frac{1}{2n}\sum_{i=1}^{n}\begin{pmatrix}(\mu_A(x_i)-\mu_B(x_i))^2 + (\nu_A(x_i)-\nu_B(x_i))^2 \\ +(\pi_A(x_i)-\pi_B(x_i))^2\end{pmatrix}} \quad (7)$$

where $d(A,B)$ is the normalized Euclidean distance between cluster A and B.

3) Combine the two closest clusters into a bigger cluster.

4) Compute dissimilarities between the new cluster and the other clusters whilst the remaining dissimilarities remain unchanged.

Several types of measurement are suitable for measuring the distance between clusters. As a widely used measure, the mean distance, is used in IFAHC:

$$d_{mean}(\mathrm{C}_j, \mathrm{C}_k) = |m_j - m_k| \quad (8)$$

where m_j is the mean value of cluster C_j.

5) Repeat Steps 3 and 4 until all objects are in one cluster or user defined termination condition is satisfied.

III. Music Recommendation: A Hypothetical Application

Suppose that 10 heavy metal music pieces have been randomly chosen from a folksonomy. Four types of heavy metal music genres have been chosen to be the attributes of the sample datasets in this study. The four genres are Folk Metal, Symphonic Power Metal, Melodic Death Metal and Traditional Heavy Metal. The tags of each music piece are represented using IFSs as presented in Table I. The calculation steps, using IFAHC, are as follows.

A. Object representation in IFS

Assume that the music piece ID 1 was tagged by 10 users, 5 users among them tagged ID 1 as “Folk Metal”, 3 users did not tag ID 1 as “Folk Metal”, and 2 users tagged ID 1 as the other tags including “Folk”, but not “Folk Metal”.

The tags of each music piece can be represented by IFSs. The membership degree of ID 1 for the “Folk Metal” attribute is 0.5, the non-membership degree is 0.3, and the hesitancy degree is 0.2. All the IFSs are computed by Eqs. 4-6 and shown in Table I.

TABLE I. A DATA SET OF 10 MUSIC PIECES REPRESENTED BY IFSS

Id	Folk			Symphonic Power			Melodic Death			Traditional Heavy		
	u	*v*	*π*	*u*	*v*	*π*	*u*	*v*	*π*	*u*	*v*	*π*
1	0.5	0.3	0.2	0.8	0.1	0.1	0.0	1.0	0.0	0.2	0.6	0.2
2	1.0	0.0	0.0	0.6	0.2	0.2	0.1	0.7	0.2	0.5	0.2	0.3
3	0.6	0.1	0.3	0.8	0.0	0.2	0.5	0.2	0.3	0.0	1.0	0.0
4	0.7	0.2	0.1	0.2	0.8	0.0	0.6	0.1	0.3	0.1	0.5	0.4
5	0.2	0.8	0.0	0.6	0.2	0.2	0.0	0.9	0.1	0.9	0.0	0.1
6	0.9	0.1	0.0	0.7	0.3	0.0	0.2	0.6	0.2	0.5	0.5	0.0
7	1.0	0.0	0.0	0.0	0.9	0.1	0.7	0.3	0.0	0.8	0.1	0.1
8	0.5	0.4	0.1	0.2	0.6	0.2	0.1	0.8	0.1	0.0	0.9	0.1
9	0.3	0.5	0.2	0.8	0.1	0.1	0.6	0.1	0.3	0.7	0.3	0.0
10	0.1	0.8	0.1	0.1	0.6	0.3	1.0	0.0	0.0	0.2	0.6	0.2

B. *Objects clustering*

The R language was used to implement the proposed IFAHC algorithm. Firstly, the dataset of 10 music pieces was initialized as ten individual clusters: {1}, {2}, {3}, {4}, {5}, {6}, {7}, {8}, {9}, and {10}. The dissimilarities between each cluster are calculated by Eq. 7. The dissimilarity matrix is shown in Table II.

TABLE II. DISSIMILARITIES MATRIX OF THE MUSIC PIECE DATA SET

	{1}	**{2}**	**{3}**	**{4}**	**{5}**	**{6}**	**{7}**	**{8}**	**{9}**	**{10}**
{1}	0	0.163	0.205	0.264	0.205	0.147	0.317	0.158	0.236	0.317
{2}		0	0.237	0.222	0.223	0.094	0.218	0.219	0.225	0.331
{3}			0	0.215	0.332	0.185	0.319	0.210	0.203	0.263
{4}				0	0.318	0.201	0.186	0.197	0.226	0.192
{5}					0	0.225	0.305	0.266	0.211	0.316
{6}						0	0.215	0.186	0.188	0.299
{7}							0	0.281	0.260	0.275
{8}								0	0.270	0.250
{9}									0	0.218
{10}										0

According to Table II, the closest two clusters are {2} and {6}. Therefore the two clusters are combined into a bigger cluster {2, 6}. The mean value of cluster {2, 6} is computed by Eq. 8 and shown in Table III.

TABLE III. IFSS OF CLUSTER {2, 6}

Folk			Symphonic Power			Melodic Death			Traditional Heavy		
u	*v*	*π*	*u*	*v*	*π*	*u*	*v*	*π*	*u*	*v*	*π*
0.95	0.05	0.00	0.65	0.25	0.10	0.15	0.65	0.2	0.50	0.35	0.15

The output of the first iteration combination is: {1}, {2, 6}, {3}, {4}, {5}, {7}, {8}, {9}, {10}. The next eight outputs for each loop are shown as follows.

- Loop 2:{1}, {2, 6}, {3, 7}, {4}, {5}, {8}, {9}, {10}
- Loop 3:{1}, {2, 6}, {3, 7}, {4, 8}, {5}, {9}, {10}
- Loop 4:{1}, {2, 6}, {3, 7}, {4, 8}, {5, 9}, {10}
- Loop 5:{1}, {2, 6, 10}, {3, 7}, {4, 8}, {5, 9}
- Loop 6:{1}, {2, 6, 10, 3, 7}, {4, 8}, {5, 9}
- Loop 7:{1}, {2, 6, 10, 3, 7, 4, 8}, {5, 9}
- Loop 8:{2, 6, 10, 3, 7, 4, 8, 1}, {5, 9}
- Loop 9:{2, 6, 10, 3, 7, 4, 8, 1, 5, 9}

After 9 iterations, all music pieces are in a big cluster. The result can be illustrated in the form of a dengrogram, as shown in Fig. 1.

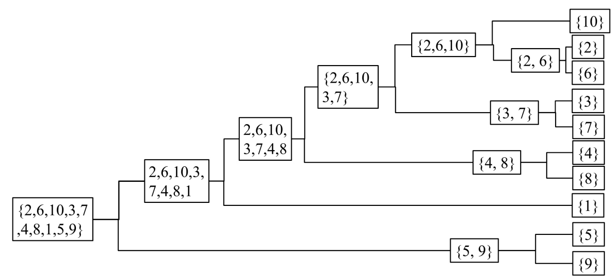

Fig. 1. Dengrogram of sample music dataset produced by IFAHC

The clustering results and dengrogram can be used to recommend music pieces to users. For example, the clustering results at Iteration 6 suggest that the data set should be separated into four clusters, {1}, {2, 6, 10, 3, 7}, {4, 8}, and {5, 9}. Assume a user listened to music piece 3, the system will recommend music pieces in order of 7, 2, 6, 10 to the user from the bottom-up direction of the dengrogram.

IV. CONCLUSIONS

This paper proposes the Intuitionistic Fuzzy Agglomerative Hierarchical Clustering (IFAHC) algorithm which was developed on the basis of the Intuitionistic Fuzzy Set (IFS) concept and Agglomerative Hierarchical Clustering (AHC). The tag objects are represented by IFS and divided into clusters using IFAHC to produce a dengrogram which can then be used for recommendation purposes. A music recommendation application of social tagging is used to demonstrate the usability and validity of the proposed approach.

REFERENCES

[1] AK. Jain, MN. Murty and P J. Flynn, “Data clustering: a review,” ACM computing surveys (CSUR), vol. 31, 1999, pp. 264-323.

[2] J. Han and M.Kamber, Data Mining: Concepts and Techniques, Southeast Asia ed., Morgan kaufmann, 2006.

[3] JH. Ward Jr, “Hierarchical grouping to optimize an objective function,” Journal of the American statistical association, 1963, vol. 58, pp. 236-244.

[4] G. Karypis, EH. Han and V. Kumar, “Chameleon: Hierarchical clustering using dynamic modeling,” Computer, 1999, vol. 32, pp. 68-75.

[5] Y. Zhao and G. Karypis, “Evaluation of hierarchical clustering algorithms for document datasets,” Proceedings of the eleventh

international conference on Information and knowledge managemen, ACM, 2002, pp.515-524.

[6] A. Shepitsen, J. Gemmell, B. Mobasher and R.Burke, “Personalized recommendation in social tagging systems using hierarchical clustering,” Proceedings of the 2008 ACM conference on Recommender systems, ACM, 2008, pp: 259-266.

[7] TV. Wal, Folksonomy coinage and definition , 2007.

[8] S. Niwa and S. Honiden, “Web page recommender system based on folksonomy mining for ITNG'06 submissions,” Information Technology: New Generations, 2006. ITNG 2006. Third International Conference on. IEEE, 2006, pp. 388-393.

[9] L. Specia and E. Motta, “Integrating folksonomies with the semantic web. The semantic web: research and applications,” Springer Berlin Heidelberg, 2007, pp. 624-639.

[10] E. Portmann, “A fuzzy grassroots ontology for improving social semantic web search,” Proceedings of 6th international summer school on aggregation operators, Benevento, 2011.

[11] K. Atanassov, “Intuitionistic fuzzy sets,” Fuzzy Sets and Systems, vol. 20 , 1986), pp. 87-96.

[12] L.A. Zadeh, “Fuzzy sets,” Inform. and Control, vol. 8, 1965, pp. 338-353.

[13] G. Klir and B.Yuan, Fuzzy sets and fuzzy logic. New Jersey: Prentice Hall, 1995.

[14] E. Szmidt and J. Kacprzyk. “Distances between intuitionistic fuzzy sets,” Fuzzy sets and systems, vol. 114(3), 2000, pp. 505-518.

[15] F. Murtagh, “A survey of recent advances in hierarchical clustering algorithms,” The Computer Journal, 1983, vol. 26, pp. 354-359.