ISSN 2007-9737

Survey of Overlapping Clustering Algorithms

Beatriz Beltrán, Darnes Vilariño

Benemérita Universidad Autónoma de Puebla,
Faculty of Computer Science,
Mexico

{bbeltran,darnes}@cs.buap.mx

Abstract. This paper is presented as a study of the overlapping clustering algorithms that have been developed in the last years, researchers have been working on these algorithms in different ways, in some cases they are based on widely known algorithms such as k-means and in others that work with heuristics or graphs. The need to work in clustering algorithms with overlap is due to the fact that currently there are many problems that require that the obtained groups be non-exclusive and for which it gives the guideline for this analysis. The algorithms included in this analysis are: ADditive CLUstering, Overlapping K-means, Dynamic Overlapping Clustering based on Relevance, Overlapping Clustering based on Density and Compactness, MCLC, A tree-based incremental overlapping clustering method, INDCLUS and Hybrid K-means.

Keywords. Clustering algorithms, supervised classification, overlapping cluster.

1 Introduction

The classification of objects or elements according to the similarities is one of the fundamental bases to learn and understand. The classification of elements arises in the human being since childhood, for example, to place objects by colors or shapes. The cluster analysis helps in the development of methods and algorithms to group and classify. Also, the problem of clustering data is being widely studied in data mining and machine learning, being its applications included to sumaries, learning, image segmentation, and marketing.

There are different ways to classify clustering algorithms, in particular, by the type of obtained clusters, [14] propose the following classification: **disjoint**, when an element belongs to exactly one cluster, for example, cluster movies by their content(AA, A, B, B15, C and D), in **fuzzy**, when an element belongs to all clusters but with a certain degree of belonging, for example, the clustering of a range of a million colors; and finally those that are **overlapped**, where an element may belong to more than one cluster, for example, the likes of feeding people.

Another categorization, where an exclusive and non-exclusive classification is proposed, is indicated in [14]. The first considers disjoint clusters, and the second one allows overlaps. Within the exclusive classification, it is used the intrinsic, where a proximity matrix is used, it is also known as unsupervised learning. The extrinsic classification uses labels for the elements.

The intrinsic classification is sub-classified in hierarchical and partitioned depending on the imposed structure of the data. The hierarchical clustering can be divisive (these algorithms form clusters by separating the existing ones) considering some similarity measure. The partitioned clustering takes into account a k parameter, which indicates the number of clusters. This taxonomy is shown in 1.

Following with the hierarchical algorithms, different authors have developed algorithms of this kind, working on different domains [7, 17], with coverage subgraphs to cluster documents [2], density subgraphs [3], suffix trees [22], center based [9], density [11], objective functions and dendograms [12], using the closest neighbor for the

doi: 10.13053/CyS-24-2-3391

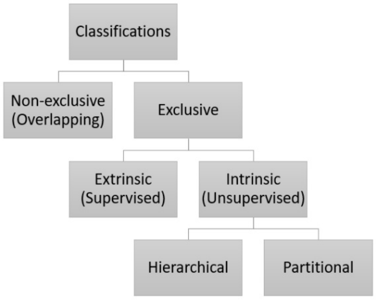

Fig. 1. Classification types (referenced by *Jain & Dubes*)

identification of duplicates [5] and predictions [10], among others.

Among the iterative algorithms, there are jobs where researchers work with maximum likelihood [16], using EM with Gaussian Mixtures [21], hybrid algorithms using GSA y K-means [13], etc. The used techniques are varied and with different results, these techniques have been applied in different types of data, such as texts, images and with discrete, numeric and categorical data.

In the particular case of the present study, it is carried out on clustering algorithms with overlap. The work is organized as follow, it starts with an explanation of the chosen algorithms, to continue in the next section with an analysis of the computational behavior of the presented algorithms, and the specification of the tested data with each technique, the last section will contain the conclusions.

2 Clustering Algorithms with Overlap

This section explains the solutions that different authors have given to the clustering problem, taking into account overlapping clusters.

2.1 ADditive CLUStering (ADCLUS)

One of the first works is [1] where a new model of clustering is described, in this model, the restrictions of the clustered objects in an exhaustive or mutually exclusive categories is relaxed allowing the establishment of overlapping clusters. It is notorious that many datasets to be grouped do not require exclusive clusters, it is from there the need of a solution with overlap, but create all possible overlapping sets gives a total of 2^{n-1} clusters, therefore, heuristics are needed to select potential groups.

ADCLUS is the proposed model, which cluster elements that meet some property, with a certain weight. ADCLUS considers *n* objects to be grouped and a symmetric proximity matrix $M = n(n-1)/2$. The data is transformed into similarities in the $[0,1]$ interval. The underlying basic equation of ADCLUS is:

$$\hat{S}_{ij} = \sum_{k=1}^{m} w_k P_{ik} P_{jk}, \tag{1}$$

where $\hat{S}_{ij}$ is theoretically the reconstructed similarity in the objects between i y j, w_k is a non-negative weight. So, the similarity between a pair of objects is the sum of the weight of those groups that contain both objects. In addition, the MAPCLUS algorithm builds a matrix starting of a set of weights and defined subsets, surpassing ADCLUS.

2.2 Overlapped K-Means (OKM)

The reason for performing an overlapping algorithm based on K-Means[4], is due to different applications in information retrieval, natural language processing, chemistry, biology, medicine; among others, where an overlapping data coverage is required, so an objective criterion is proposed associated with the OKM algorithm that generalizes the K-means algorithm.

The objective criterion is defined as follows: Given a set of data vectors $\mathcal{X} = \{x_i\}_{i=1}^{n}$ with $x_i \in \mathbb{R}^n$, to find a *k-way* coverage $\{\pi_c\}_{c=1}^{k}$, where π_c represents the c^{th} group; such that the following goal is minimized:

$$\mathcal{J}(\{\pi_c\}_{c=1}^{k}) = \sum_{x_i \in \mathcal{X}} \|x_i - \phi(x_i)\|^2, \tag{2}$$

where each x_i must be in at least one group, so: $\bigcup_{c=1}^{k} \pi_c = \mathcal{X}$ y $\phi(x_i)$ denotes the image of x_i defined by the combinations of the (m_c) prototypes for the group x_i as:

$$\phi(x_i) = \frac{\sum_{A_i} m_c}{|A_i|}, \tag{3}$$

where A_i the assignment set of: $x_i : \{m_c | x_i \in \pi_c\}$. So an heuristic for obtaining the optimal coverage was developed.

2.3 Dynamic Overlapping Algorithm based on Relevance (DClustR)

The DClustR algorithm [18] is an algorithm that allows the overlap between its groups, as an alternative for analysis in social networks, information retrieval and bioinformatics, this algorithm is based on graph theory by introducing strategies for the construction of more precise overlapping clusters or when the collection changes.

The main idea is to generate a set of clusters that are a coverage $\tilde{G}_\beta$ using *ws-graphs* and subsequently improve the initial clusters to obtain the right one, so “improve” means reducing both the number of clusters as the overlap between them.

To define the *ws-graph*, let $G_\beta = \langle V^*, E^*, S \rangle$ a graph of similarity threshold with weight. A star shape weighted subgraph (*ws-graph*) in$\tilde{G}_\beta$, denoted by $G^*_\beta = \langle V^*, E^*, S \rangle$ is a subgraph of $\tilde{G}_\beta$ having a vertex $c \in v^*$ such that there is an arch between c and another vertex un V^*. The vertex c is called the center of the *ws-graph* and the rest of the vertexes are called *satellites*. Isolated vertexes are considered degenerate *ws-graphs*.

Being the *ws-graph* determined by its center, the problem is to build the set $W = \{G^*_{c_1}, G^*_{c_2}, ..., G^*_{c_k}\}$ of *ws-graphs*, such that W if a coverage of $\tilde{G}_\beta$, can be seen as the problem of reconstructing the set $X = \{c_1, c_2, ..., c_k\}$ such that $c_i \in X$ is the center of $\tilde{G}_{c_i} \in W, \forall i = 1, ..., k$.

To avoid analyzing all vertexes in v^* and delimit the search space, a selection criterion is established, DClustR introduces the concept of relevance of a vertex, for which it can be selected those vertexes with the highest degree, and its about maximizing the number of added vertexes to the $\tilde{G}_\beta$ coverage in each iteration.

2.4 Overlapping Clustering based on Density and Compactness (OCDC)

The OCDC algorithm [19], introduces a new graph coverage and a new filter strategy, with which a small set of overlapping clusters are can be obtained. The collection of objects are represented as a graph of similarity threshold with weight $\tilde{G}_\beta$, the overlapped clustering is realized in two phases: the *initialization* and the *improvement*.

In the *initialization* phase, an initial set of groups of coverage vertexes is built $\tilde{G}_\beta$, using the *ws-graphs*, in this context, each of these graphs make up an initial group. In this step, the algorithm seeks to reduce the search space, and OCDC introduces the concepts of *density* and *compactness* of a vertex v. The *density* of a vertex $v \in V$, is calculated by using the following equation:

$$v.density = \frac{v.pre\_dens}{|v.Adj|}, \tag{4}$$

where $v.pre\_dens$ is the number of adjacent vertexes to v, having a degree not greater than the degree of v, and $v.Adj$ is the total of adjacent vertexes to v.

The compactness of a vertex $v \in V$, is estimated as follow:

$$v.compactness = \frac{v.pre\_compt}{v.Adj}, \tag{5}$$

where $v.pre\_compt$ is the number of vertexes $u \in v.Adj$ such that $Aprox\_Intra\_sim(G^*_v)$
$\geq Aprox\_Intra\_sim(G^*_u)$, where G^*_v and G^*_u are *ws-graphs* determined by v y u, respectively. The greater value of the compactness is included in coverage and therefore it is the best coverage graph.

2.5 MCLC Algorithm

The MCLC algorithm is proposed to discover overlapping communities [6], for which is used a random path in a line graph and attraction intensity. Unlike the traditional random path that starts from a node, it starts from a link. In the first instance, a network graph is transformed into a weighted linear graph, and the random path in this linear graph is associated with a string of Markov. In order to obtain the probability of the Markov's chains, a similarity between the pair of leagues is obtained. Then, the leagues can be grouped into "league communities" where those nodes can be overlapped.

The *league communities* become "node communities", and a attraction intensity is defined for the control of the overlap's size. Finally the communities that allow overlapping are detected.

The distance or similarity between pairs of leagues is obtained by calculating the probability transition of random paths in the linear graph. A matrix of $M \times M$, can be associated to a chain of M-Markov's states, the transition matrix $P = [P_{\alpha\beta}]$, is defined as:

$$P_{\alpha\beta} = \frac{h_{\alpha\beta}}{\sum_{\beta} h_{\alpha\beta}}. \qquad (6)$$

The number of repetitions of random paths that start from the league α, $[P^t]_{\alpha\beta}$ should be considered and is the probability that the path starts from α and remain in β which is $\sum_{t=1}^{T}[P^t]_{\alpha\beta}$ $(1 \geq t \geq T)$.

The cluster analysis can use "peer league pairs" in communities of candidate leagues, so a similarity (symetric)$\phi_{\alpha\beta}$ is proposed as follow:

$$\phi_{\alpha\beta} = \phi_{\beta\alpha} = \sum_{t=1}^{T}([P^t]_{\alpha\beta} + [P^t]_{\beta\alpha}). \qquad (7)$$

The distance $d_{\alpha\beta}$ between pair of leagues (α, β) is obtained by the complement of the similarity and normalization of the results between 0 and 1:

$$d_{\alpha\beta} = d_{\beta\alpha} = 1 - \frac{\phi_{\alpha\beta} - min\ \phi}{max\ \phi - min\ \phi}. \qquad (8)$$

2.6 Clustering Method with Incremental Overlapping based on Trees

The clustering method with incremental overlap based on trees [20], uses the tripartite decision theory. A tree is represented by points that can improve the relevance of the search result. The overlapped clusters are represented by the tripartite decision with a set of intervals. Tripartite decision strategies are designed for the update of clusters, at the moment that the data increases. Further, with this method is possible to determine the number of clusters during the process.

To define the tripartite decision clustering, be $U = \{x_1, ..., x_n, ..., x_N\}$ the universe, and the resulting clusters $\mathbf{C} = \{C_1, ..., C_k, ..., C_K\}$ a family of clusters of the universe. x_n is an object, which has D attributes, $x_n = (x_n^1, ...x_n^d, ..., x_n^D)$, where x_n^d represents the value of the *d*-th attribute of the object x_n, where $n \in \{1, ..., N\}$ y $d \in \{1, ..., D\}$.

The algorithm starts calculating the distance(Euclidean) between objects. The similarity between the objects is obtained with the complement of distance. Subsequently, the representative points are calculated using the following condition: if $|Neighbor(r)| \geq \zeta$, r is a representative point and represents the object in the area where r is centered and with radius δ.

The next step is the construction of an indirect G graph based on the R representation points, this is achieved using the tripartite decision and the calculated representative points. Finally, the algorithm search in the subgraph those who are strongly connected in the graph G. This design allows the growth of the data, but when increasing, it is required to simulate different situations to evaluate the performance of the method.

2.7 INDCLUS

In this section is examined the scalability of the ADCLUS and INDCLUS models [8], which are techniques that can be used to extract overlapping clusters with similar data. In this paper was taken the models ADCLUS and INDCLUS appropriately and were designed different metaheuristics extensions to have more relaxed models.

For the INDCLUS model, N elements are considered, with a similarity matrix $S = (s_{ij})_{\{N \times N\}}$, and is required a group of a known number of M overlapped clusters possibilities. The INDCLUS model requires to minimize the optimization functions:

$$
\begin{aligned}
& min \sum_{k=1}^{K} \sum_{i=1}^{N} \sum_{j \neq i} \left(S_{kij} - \sum_{m=1}^{M} w_{km} P_{im} P_{jm} - c_k \right)^2 \\
& w_{km} \geq 0, \forall k = 1, ..., K; m = 1, ..., M, \\
& c_k \geq 0, \forall k = 1, ..., K, \qquad (9) \\
& P_{im} \in \{0, 1\} \forall i = 1, ..., N; m = 1, ..., M,
\end{aligned}
$$

where K is the number of subjects, N is the number of elements to be stored and s_{kij} is the similarity of the elements i and j del sujeto k. If $K = 1$ the model is reduced to ADCLUS.

The used heuristics with these algorithms are: alternating approach of minimum squares (SINDCLUS), a symmetric approach applied to SINDCLUS (SYMPRES), simulated annealing (SA-SINDCLUS), tabu search (TABU-SINDCLUS), and relaxed solution space (SMC-Relax). The tests were realized with medium size real datasets, SMC-Relax had the a better execution than SINDCLUS and SYMPRES. The use of heuristics makes the ADCLUS and INDCLUS models scalable.

2.8 Hybrid K-Means

In [15] is described an algorithm (HKM-OKM) that combines harmonic k-means with overlapped k-means. By making use of the overlapped k-means algorithm; which is an extension of k-means is sensitive to the centroid of the initial cluster but when is combined with harmonic k-means this limitation can be overcome.

The main idea in this method is to use the output HKM method to initialize the centroids of the OKM method. The OKM method was explained in section 2.2. The HKM algorithm introduces a bias(using the weight) to move the cluster centers to the data points that are most important according some criteria.

Similar to the k-means algorithm, the HKM method can be formulated as an optimization problem where the objective is to minimize:

$$
Q''(\pi) = \sum_{i=1}^{n} \frac{k}{\sum_{j=1}^{k} \frac{1}{\|\overrightarrow{x_i} - \overrightarrow{z_j}\|^p}}, \qquad (10)
$$

where p is a free parameter (typically $p \geq 2$), and the expression $\left(\frac{k}{\sum_{j=1}^{k} \frac{1}{\|\overrightarrow{x_i} - \overrightarrow{z_j}\|^p}} \right)$ is the harmonic media. To calculate the harmonic media, the algorithms needs to calculate the cluster centroid $\overrightarrow{z_j}$ by using:

$$
\overrightarrow{z_j} = \frac{\sum_{i=1}^{n} m(\overrightarrow{z_j} | \overrightarrow{x_i}) w(\overrightarrow{x_i}) \overrightarrow{x_i}}{\sum_{i=1}^{n} m(\overrightarrow{z_j} | \overrightarrow{x_i}) w(\overrightarrow{x_i})}, \qquad (11)
$$

where $m(\overrightarrow{z_j} | \overrightarrow{x_i})$ is a member of the data point $\overrightarrow{x_i}$ to the cluster centroid j calculated by:

$$
m(\overrightarrow{z_j} | \overrightarrow{x_i}) = \frac{\|\overrightarrow{x_i} - \overrightarrow{z_j}\|^{-p-2}}{\sum_{j=1}^{k} \|\overrightarrow{x_i} - \overrightarrow{z_j}\|^{-p-2}}. \qquad (12)
$$

And $w(\overrightarrow{x_i})$ is the associated weight which each $\overrightarrow{x_i}$ point, calculated by:

$$
w(\overrightarrow{x_i}) = \frac{\sum_{j=1}^{k} \|\overrightarrow{x_i} - \overrightarrow{z_j}\|^{-p-2}}{\left(\sum_{j=1}^{k} \|\overrightarrow{x_i} - \overrightarrow{z_j}\|^{-p} \right)^2}. \qquad (13)
$$

The HKM-OKM algorithm starts by finding centers, using the HKM, initializes OKM using the found centers. A set of medical data is used, because it is required to model elements with overlap. It improves the obtained results by OKM.

3 Algorithms Analysis

The analyzed algorithms have a maximum time complexity of the quadratic order as they are: OCDC, MDLC based on trees, INDCLUS with hybrid heuristics and k-means; the particular case of OKM maintains the order of the algorithm on which it was based (k-means), and only the ADCLUS algorithm is of the cubic order. This information can be reviewed in table 1.

In the experiments the number of instances that were used with these algorithms varies, being the minimum of 105 and a maximum of 102, 294 instances. Further, the objects are of different types: discrete, qualitative or documents. In all experiments were tested stable datasets, for

Table 1. Comparison between clustering algorithms with overlap.

Algorithm	Datatype	Amount of Data	Complexity
ADCLUS	Discrete	105	$O(n^3)$
OKM	Qualitative - Documents	1,308	$O(t{\cdot}\, n{\cdot}\, k \log k)$
DClustR	Qualitative – Documents	16,006	$O(\|V\| + \|E_\beta\|)$
OCDC	Documents	16,006	$O(n^2)$
MCLC	Discrete	1,133	$O(m^2 n)$
Based on Trees	Discrete	5,473	$O(n^2 + n \log n)$
INDCLUS	Qualitative – Documents	102,294	$O(n^2)$
Hybrid K-Means	Qualitative	699	$O(n^2)$

example, the repository *UCI Machine Learning*[1] is used, testing the cancer dataset, heart disease, parkinson, among others, also the dataset on the *karate Zachary*[2], KDD, ISOLET were used; *Reuters-21578*[3], TDT2[4], were mainly used for the dataset with documents.

In general, clustering algorithms with overlap are based on others algorithms that do not support overlap and even improve some aspects of the same.

4 Conclusion and Future Work

In this article, the clustering algorithms with overlap were analyzed.

Over the last years, the interest in the development and improvement of this type of algorithms has been constant and researchers continue to seek to improve the obtained results.

Different techniques have been used in this type of algorithms, from basic algorithms such as k-means, using heuristics to have scalability. In addition, graph theory has also been used and finally, the combination of algorithms was used to counteract some deficiencies of the algorithms.

The amount of elements that have been worked with these algorithms isn't very large in general, standardized datasets are used and the quality of the algorithms is verified by standard measures such as *F-Measure* or *F-Bcubed*.

[1] https://archive.ics.uci.edu/ml/index.php

[2] http://konect.uni-koblenz.de/networks/ucidata-zachary

[3] http://www.daviddlewis.com/resources/testcollections/reuters21578/

[4] https://catalog.ldc.upenn.edu/LDC2001T57

Acknowledgements

We would like to thank to the vice-rectory of research and postgraduate studies (*VIEP*) of the Benemérita Universidad Autónoma de Puebla for the economical support.

References

1. **Arabie, P., Carroll, J. D., DeSarbo, W., & Wind, J. (1981).** Overlapping clustering: A new method for product positioning. *Journal of Marketing Research*, Vol. 18, No. 3, pp. 310–317.
2. **Aslam, J. A., Pelekhov, E., & Rus, D. (1999).** A practical clustering algorithm for static and dynamic information organization. *Proceedings of the 1999 Symposium on Discrete Algorithms*, pp. 51–60.
3. **Aslam, J. A., Pelekhov, E., & Rus, D. (2004).** The Star Clustering Algorithm For Static And Dynamic Information Organization. *Journal of Graph Algorithms and Applications*, Vol. 8, No. 1, pp. 95–129.
4. **Cleuziou, G. (2008).** An extended version of the k-means method for overlapping clustering. *2008 19th International Conference on Pattern Recognition*, pp. 1–4.
5. **Costa, G., Manco, G., & Ortale, R. (2009).** An incremental clustering scheme for data de-duplication. *Data Mining and Knowledge Discovery*, Vol. 20, No. 1, pp. 152.
6. **Deng, X., Li, G., Dong, M., & Ota, K. (2017).** Finding overlapping communities based on markov chain and link clustering. *Peer-to-Peer Networking and Applications*, Vol. 10, No. 2, pp. 411–420.

7. **Fisher, D. H. (1987).** Knowledge acquisition via incremental conceptual clustering. *Machine Learning*, Vol. 2, No. 2, pp. 139–172.

8. **France, S. L., Chen, W., & Deng, Y. (2017).** Adclus and indclus: analysis, experimentation, and meta-heuristic algorithm extensions. *Advances in Data Analysis and Classification*, Vol. 11, No. 2, pp. 371–393.

9. **Gan, G., Ma, C., & Wu, J. (2007).** *Data clustering - theory, algorithms, and applications.* SIAM.

10. **Gan, H., Fan, Y., Luo, Z., & Zhang, Q. (2018).** Local homogeneous consistent safe semi-supervised clustering. *Expert Systems with Applications*, Vol. 97, pp. 384–393.

11. **Ghosh, J., Liu, A., & Gupta, G. (2008).** Automated hierarchical density shaving: A robust automated clustering and visualization framework for large biological data sets. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol. 7, pp. 223–237.

12. **Gilpin, S. & Davidson, I. (2017).** A flexible ilp formulation for hierarchical clustering. *Artificial Intelligence*, Vol. 244, No. C, pp. 95–109.

13. **Hatamlou, A., Abdullah, S., & Nezamabadi-Pour, H. (2012).** A combined approach for clustering based on k-means and gravitational search algorithms. *Swarm and Evolutionary Computation*, Vol. 6, pp. 47–52.

14. **Jain, A. K. & Dubes, R. C. (1988).** *Algorithms for Clustering Data*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

15. **Khanmohammadi, S., Adibeig, N., & Shanehbandy, S. (2017).** An improved overlapping k-means clustering method for medical applications. *Expert Syst. Appl.*, Vol. 67, No. C, pp. 12–18.

16. **Kneser, R. & Ney, H. (1993).** Improved clustering techniques for class-based statistical language modelling. *EUROSPEECH*, ISCA.

17. **Patnaik, A. K., Bhuyan, P. K., & Rao, K. K. (2016).** Divisive analysis (diana) of hierarchical clustering and gps data for level of service criteria of urban streets. *Alexandria Engineering Journal*, Vol. 55, No. 1, pp. 407–418.

18. **Pérez-Suárez, A., Martínez-Trinidad, J. F., Carrasco-Ochoa, J. A., & Medina-Pagola, J. E. (2013).** An algorithm based on density and compactness for dynamic overlapping clustering. *Pattern Recognition*, Vol. 46, No. 11, pp. 3040–3055.

19. **Pérez-Suárez, A., Martínez-Trinidad, J. F., Carrasco-Ochoa, J. A., & Medina-Pagola, J. E. (2013).** A new overlapping clustering algorithm based on graph theory. **Batyrshin, I. & González Mendoza, M.**, editors, *Advances in Artificial Intelligence*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 61–72.

20. **Yu, H., Zhang, C., & Wang, G. (2016).** A tree-based incremental overlapping clustering method using the three-way decision theory. *Knowledge-Based Systems*, Vol. 91, pp. 189–203. Three-way Decisions and Granular Computing.

21. **Yu, J., Chaomurilige, C., & Yang, M.-S. (2018).** On convergence and parameter selection of the em and da-em algorithms for gaussian mixtures. *Pattern Recognition*, Vol. 77, pp. 188–203.

22. **Zamir, O. & Etzioni, O. (1998).** Web document clustering: A feasibility demonstration. *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR ’98, ACM, New York, NY, USA, pp. 46–54.

Article received on 30/10/2019; accepted on 12/03/2020.
Corresponding author is Beatriz Beltrán.