Information enhancement in a voluminous forum with automatic co-clustering

François Rioult, Sylvain Ferrandiz, Monique Bastien, Marc Boullé

▶ **To cite this version:**

François Rioult, Sylvain Ferrandiz, Monique Bastien, Marc Boullé. Information enhancement in a voluminous forum with automatic co-clustering. International Symposium on Web AlGorithms, Jun 2016, Deauville, France. hal-01636383

HAL Id: hal-01636383
https://hal.science/hal-01636383

Submitted on 16 Nov 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Information enhancement in a voluminous forum with automatic co-clustering

François Rioult
University of Caen - Normandy
francois.rioult@unicaen.fr

Sylvain Ferrandiz
PredicSis
sylvain.ferrandiz@predicsis.com

Monique Bastien
Caméra Vidéo

Marc Boullé
Orange Labs
marc.boulle@orange.com

Abstract

Any web forum is a typical example of a structured source of textual data. Maintaining such a forum requires many manual processing from the webmaster, even more as the volume of data grows. The quality of the structure and the moderation process is thus strongly dependent on the available time. In this article, we are interested in applying data mining tools, especially co-clustering algorithms, in order to propose some automation. We illustrate and validate our proposal on cameravideo.net's forum[1]*. The website cameravideo.net was the premier french website dealing with amateur and professional video.*

I. Introduction

On the Internet, forums allow users to exchange information about their favorite subjects, such as video (hardware, shooting, editing, etc.). Well-known usages are related to questions/answers: information gathering, support, contact. These forums are also useful to deal with users' contributions, tutorials and various articles. They lastly establish a place for a social community sharing a common interest by proposing additional services such as mails and feeds, that stimulate exchanges.

Although forum softwares are efficient for dealing with users and content, information management and enhancement are not taken into account. Firstly, the forum layout is often rigid while a member may like to post a comment in several categories. Secondly, search in a forum is limited to basic text matching, even if the writer may tag his/her contribution in order to promote its access. There are methods that automatically index the content but enhancing this index to optimize its structure requires a colossal human effort. The daily management is also tedious and costly because it requires a lot of moderation.

Our aim is to study how data mining techniques can lead management tasks to automation, in order to better present and enhance information in forums. We more precisely focus on applying *co-clustering methods* to text content in a forum, improving the flexibility of its structure and making its daily management easier. We are interested in building a process that evaluates the relevancy of the used models, focused on the forum of the website `cameravideo.net`. In particular, we study the impact of filtering and lemmatizing the vocabulary.

Co-clustering will provide groups of words that could allow to improve the forum structure. Different categories could then be suggested to the user, helping the new content to be indexed in the right place. Co-clustering will also group texts, a new group allowing to detect new trends and to affect the structure of the forum.

The article is organized as follows: in Section II, we describe a methodology for enhancing and assessing the information in the textual data. Enhancing information relies on using a co-clustering algorithm and assessing relies on a supervised analysis. We review the state of the art and justify our choices. In Section III, we describe the data, coming from the forum `cameravideo.net`, on which we applied our methodology. In Section IV, we put the methodology to the test through an unsupervised analysis, a descriptive analysis then a supervised analysis of the forum data.

II. Method and statistical tools

Computing the vocabulary is the first stage of natural language processing with bag-of-words approach. There are different strategies and we do not want to build a new one but rather develop a non parametric and non agnostic methodology for enhancing the information. We here describe this methodology and the statistical tools we used.

I. Methodology

We proposed the following methodology:

1. Define the textual units, for example the thread title, the title of the first post, etc.
2. Define a vocabulary and get an occurrence matrix *textual unit* × *words of the vocabulary*.
3. Evaluate the correlation between the texts and the words with the help of a co-clustering algorithm.
4. Validate the groups of texts with projecting a target variable.

The first two steps consisted in transforming the raw textual data into a data table from which a sta-

[1] Since the beginning of this study, one of its author (Monique Bastien) passed away. She was the webmaster of the website `http://www.cameravideo.net`, that also disappeared, though it is still accessible with the web archive, see for example `https://web.archive.org/web/20121115053234/http://www.cameravideo.net/`.

tistical analysis could be handled. Different strategies may lead to extract the vocabulary into different tables. For example, a lemmatization task could be used or not.

For each representation stemming from this choice, we led a statistical analysis in order to assess its relevancy. This analysis was based upon applying a co-clustering algorithm (Step 3). This algorithm should allow the methodology to be non parametric: in particular, the number of groups of each partition should be an output of the algorithm and not an input.

In order to complete this non-parametric assessment, we also propose to use a supervised assessment (Step 4). It estimates the results of the non supervised one through a correlation between the groups and a target variable which is explicable by the user. Moreover, it takes advantage of a usual and easily explicable criterion: the empirical risk.

II. Co-clustering algorithms

Co-clustering algorithms have been brought up to date by the need for analyzing large binary matrices, for example coming from consumers' bags, textual corpus, gene expression data, etc. Such algorithms compute a partition of the rows and a partition of the columns[2] of the matrix. It is now a usual unsupervised classification technique in statistics and data mining.

Our strategy consisted in first applying a classical clustering algorithm on the rows and then on the columns [12] with the information bottleneck principle (the best trade-off between accuracy and complexity) and the clustering algorithm stemming from it. This kind of approach builds a conditional analysis.

The works of [8] and [9] led to co-clustering algorithms allowing a joint analysis of both the columns and the rows of the matrix. Such method lies on considering a global evaluation criterion, that compares the quality of two couples of partitions. The method also optimizes this criterion using a heuristic. In [8], a cutting criterion for bipartite graphs is minimized and the algorithmic solution searches for singular vectors. In [9], the criteria is the Kullback-Leibler divergence between the initial joint density and the one obtained after partitioning the rows and columns. Analyzing the criteria led to an iterative algorithm that alternatively optimizes the partition. An approach generalizing to the Bregman divergences was studied in [1].

These algorithms however require the user to set the number of groups in the partitions. That is why we focused on the `Khiops` algorithm, described in [5], whose theoretical background in based upon in [4]. This method has proved to be efficient during a challenge [5] and provides a probabilistic criteria measuring the ability of the model to explain the data. This measure is very useful for comparing different models.

[2] Bi-clustering computes associations between a group of rows and a group of columns.

III. Models for supervised classification

There are a lot of data analysis softwares. We chose to use `Khiops`[3] [10] from the Orange Labs. It lies on the MODL approach [4], that leads to non parametric (no assumption about the underlying distribution) learning algorithms that are time efficient and provide reliable models. It is also based upon [3] for selecting the variables and averaging the models.

The supervised model uses a naive Bayes algorithm which is improved by the three following points:

1. The single-variable conditional probabilities are estimated by MODL algorithms for discretizing the numerical variables and grouping the categorical ones.
2. The descriptive variables are selected through the MODL approach.
3. The models are weighted by a Bayesian averaging scheme.

The modeling method is efficient in terms of the predictive performance of the built models. On several different challenges, the method indeed frequently produced the best predictive model.

The theoretical complexity of `Khiops` is $O(N\sqrt{N}\log(N))$ [6] where N is the number of instances in the database. In a corpus of texts over W words, the instances are the pairs $(text, word)$ then N if the number of use of words in the texts. For example, on 20,000 texts over 10,000 words, $N = 20,000,000$. In practice, on a modern computer, a few hours are necessary for each billion of instances. It is quite long, but this analysis has not to be done every day.

III. Description of the data

The forum of the website `cameravideo.net` was organized with categories, each one consisted of threads. Each thread consisted of comments that were posted by the users.

There were 49 categories, for example "video news", "rent, buy, sell", "Sony Vegas - DVD architect", "Apple - Mac software". They structured the forum and the moderators *a priori* decided on them. Through long term observation, the moderators may consider modifying this structure, by adding new categories or by merging some of them.

We had 8,528 threads. They are non uniformly distributed among the categories; for example, "video news" has 1,303 threads but "Apple -

[3] http://www.khiops.predicsis.com/

Mac software" has only two threads. Each thread had a title, such as "Welcome on the forum" or "Which Super-8 camera should be chosen?". Lastly, there were 61,044 comments posted on these threads.

Bottleneck The moderation task consists in validating the subject of a thread, potentially replacing the thread in the dedicated category and making sure the comments fit the policy of the site and stay in the perimeter of the thread. Facing the volume of the contributions, many moderators are needed.

Tools for vocabulary retrieval Even if the relation between terms was a promising aim for information retrieval, we did only simple processing of the text and stayed very far from complex approaches such as ontologies [2, 11]. Our goal was namely related to the design of a simple, effective and not time-consuming method, without resources such as dictionaries, taxonomies or ontologies.

The vocabulary was defined as follows: a *word* is a string of non-space characters delimited by two spaces. One-letter words were removed. The lemmatization was optional but could be carried out by the `treetagger`[4]: it needed only a few minutes for the whole corpus.

IV. Information enhancement through co-clustering: experiments

In this section, we show that a co-clustering algorithm is an operational way to the enhancement of information in voluminous forums. We analyzed the results of this algorithm in order to validate the choices regarding the pre-treatment stage (definition of the textual unit, filter the words according to their size or frequency, lemmatization). To this aim, we first used an unsupervised approach then a supervised one.

I. Unsupervised analysis

In order to study the possibilities of automatically structuring and moderating the forum, we asked the following question, illustrating the interest of an unsupervised analysis:

Question 1. – Were the titles of the thread enough for categorizing the threads?

To answer this question, we prepared the raw data as follows: the textual unit was the title of the thread; we got a vocabulary of 12,902 words, thus the occurrence matrix with 8,528 rows (one row each textual unit) and 12,902 columns (one column each word in the vocabulary).

We applied the co-clustering algorithm on this matrix in order to obtain a partition of the rows and a partition of the columns. The underlying MODL approach allows the algorithm to consider and compare partitions with different sizes: the user does not have to set the number of groups of words or texts *a priori*. In this case, the solution proposed by the algorithm had only one single group of words and one single group of texts. We then rejected the hypothesis that the words of the title could categorize the threads. On a statistical point of view, the words of the titles did not bring any information about the thread.

From this analysis, we considered the first comment of each thread and asked the question:

Question 2. – Did the first comment of each thread allow to compute a categorization for the threads?

To answer this question, we considered the textual unit as the concatenation of the title and the first comment. The matrix had 8,528 rows and 53,119 columns and we applied the co-clustering algorithm. This time, we got 278 groups of words and 214 groups of threads. The first comment then brought enough information for the threads to be categorized.

Note that we have no clue about the quality of these groups, this will be further discussed in the descriptive analysis (next section).

We then asked the last question:

Question 3. – What was the contribution of the lemmatization for defining the vocabulary?

For the titles and the first comments of each thread, we computed a lemmatized vocabulary. The matrix had 8,528 rows and 41,582 columns. The co-clustering algorithm provided 282 groups of words and 204 groups of threads.

We used the *a posteriori* probability of each model, knowing the data D, for comparing the two co-clustering M_1 and M_2: $p(M_1/D)$ and $p(M_2/D)$. The co-clustering namely optimizes this criteria and returns the couple of partitions whose *a posteriori* probability is the highest. A linear transformation of this criteria provides an explicable indicator: the compression gain. It is null when the co-clustering gives only one group for the rows and columns, it tends to 1 when the correlation is higher and the occurrence matrix is near to a diagonal one. In our situation, the compression gain for the raw vocabulary was 0.0229 and 0.0233 for the lemmatized one.

We could then conclude:

1. The first comment allows to categorize the thread.
2. Using the lemmatization does not provide any significantly better result.

This conclusion was confirmed when detailing the groups of threads or words. Such analysis is yet tedious. One more efficient way to interpret the group consisted in leading a supervised analysis. The results are given below.

II. Descriptive analysis

For the raw vocabulary (no lemmatization). – When applied to the occurrence matrix of the raw vocabulary from the title and the first comment, the co-clustering algorithm partitioned the vocabulary into 278 groups and the threads in 214 groups. Figure 1 shows the distribution of the relative size of each group.

Some groups of words covered a typical theme, that could be interpreted without knowing the threads. For example, 6 groups had only one word: {on}, {is}, {in}, {and}, {the}, {of}. One group had three similar words:

[4] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

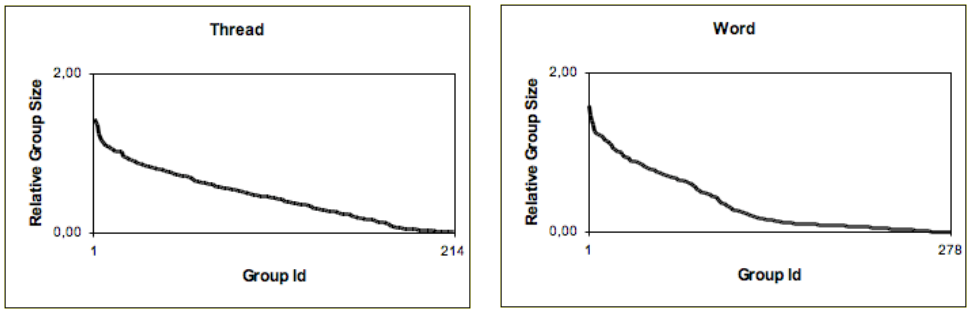

Figure 1: *Distribution of the sizes of the groups of words or threads, regarding the total number of words or threads. These sizes are from 1 to 846 for the words, from 1 to 122 for the threads.*

{#8217, #8230, #339 }. One group had 266 words containing {90, 44, street, 35, 44, fax, . . . , tel:, avenue, fax:, province, mail:, 92414, . . . }, used for writing postal coordinates or telephone numbers. One group of 185 words contained only English words[5]. One group of 131 words contained {buy, advice, expensive, budget, think, old, how much, good deal, purchase, provide, consider} that are linked to the buying theme. One group had 141 words related to vehicle (brand names). Most of the groups were of course related to the video theme of the forum: broadcasting, system (linux, PC, Mac), encoding, software, etc.

These groups of words were useful to improve the forum structure. On one side, according to the vocabulary detected while the writing of the first comment, different categories could be proposed to the user, so that the new content is indexed in the right place. On the other side, the emergence of a new group could allow to detect new trends and to affect the structure of the forum. The groups of words could also help to improve the human interface of the search functions.

About the threads, we found one group of two threads whose first comment were identical and written in English. They related to a single user promoting electronic accessories (pocket PC, cameras, PDA, smartphone) and only the titles differed: "For Sell::apple iPhone 8gb—$300usd/nokia N95 8gb(black)—-$400usd" and "For Sell::apple iPhone 48gb—$230usd/apple 16gb Ipod Touch —$190usd". One group gathered two very similar threads with redundant titles: "Canon XH A1 and G1 (HDV, 1080i, 24p) for replacing the XM2" and "the new Canon HDV HX A1 and G1 replace the XM2". This example showed how co-clustering could detect unwanted comments. Another use is related to the improvement of the forum structure, in order to present groups of homogeneous contributions, leading to a point of view which is different from the static structure of the categories.

Refining the analysis of the groups of threads would require a human effort and the design of interfaces. We however give details about the supervised evaluation of our results in the next section.

For the lemmatized vocabulary. – When applied to the occurrence matrix of the lemmatized vocabulary from the title and the first comment, the co-clustering algorithm partitioned the vocabulary into 282 groups and the threads in 204 groups. Figure 2 shows the distribution of the relative size of each group.

Some groups of words were explicable. For example, five groups had only one word: {on}, {in}, {of}, {the}. One group of 247 words contained {channel, TNT, satellite, decoder, TF1, ADSL, M6, DVB, receiving, VOD, box, TVHD, . . . }. This group covered the television theme. One group of 199 english words contained {home, equity, href, loan, used, buy, Man, mitsubichi, yong, ssang, porsche, Fiat, nissan, alfa, Mercedes, chrysler, seat, fuel, efficient, highly, cheap, mortgage, approach, prices, energy, viagra, phentermine, cell, didrex, zoloft, xenical, prozac . . . }, related to unwanted content or spam.

For the threads, we again found the group of two threads written by the electronic retailer. Another retailer for SIM-unlocked iPhones was also detected.

Manual analysis of the difference which are induced by the lemmatization generally did not reveal any advantage of this treatment. In both cases, the groups of words were consistent, easily explicable and easy to enhance. There were only slight differences when redundancy appeared, for example between "camera" and "cameras". Lemmatization could then simply be applied to improve the presentation of the results.

III. Supervised analysis

We led a supervised analysis of the previous results. Its principle [7] is based upon estimating the quality of the co-clustering by measuring the improvement on a supervised classification, when the group values (provided by the co-clustering) were added as a new variable. For the supervised classification, we chose the category which the thread was attached to as the target variable.

This analysis was automatic and provided a numerical evaluation based on a classical indicator: the empirical risk. It was less tedious than a descriptive analysis of the groups.

Building the classifier. – The previous unsupervised analysis allowed to design a new descriptive variable for the threads: the index of the group which they belonged to. In fact we had two such variables, one related to the raw vocabulary and the other related to the lemmatized vocabulary. To both variables, we associated a single

[5] The language for the `cameravideo.net` forum is French.

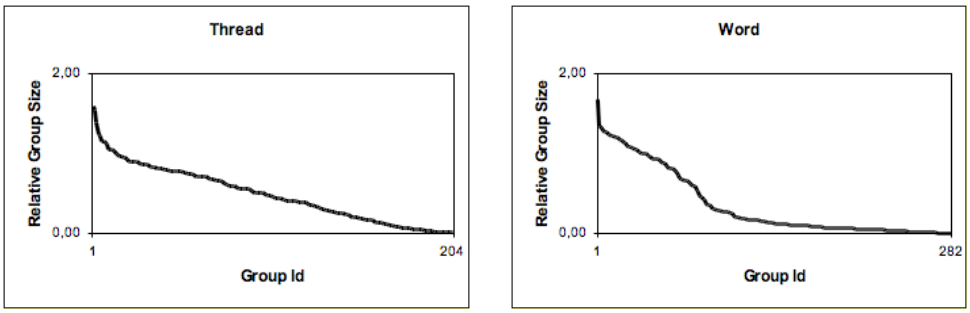

Figure 2: *Distribution of the sizes of the groups of words or threads, regarding the total number of words or threads. These sizes are from 1 to 698 for the words, from 1 to 136 for the threads.*

variable classifier, that assigns a thread to the tag that was majority in the group of threads.

We split the threads into two disjoint sets for learning and testing. The test performances of the classifiers are detailed in Table 1. We can observe that the impact of the lemmatized vocabulary is light: only 1.2%.

classifier	empirical risk
Majority	84.95%
Raw vocabulary	56.83%
Lemmatized vocabulary	55.66%

Table 1: *Empirical risk while testing the majority and univaried classifiers.*

Projecting the target variable on the thread groups allowed a direct interpretation. For example, in the case of the lemmatized vocabulary, the threads of the group #57 belonged to the category "Welcome in our forum!" for 60% and to the category "Meeting the team" for 23%. The threads of the group #144 belonged to the category "rent, buy, sell" for 94%.

Supervised modeling. – We led a second supervised experiment: we applied a supervised classification algorithm on the bag-of-words representation of the threads. This experiment allowed to give a reference performance for predicting the category of a thread, which we compared to the one obtained by adding the variable generated by the clustering.

For each vocabulary, we removed the words whose frequency was below a k threshold. Each word then allowed to build a descriptive variable: the frequency of the word in the title and the first comment. The `Khiops` tools was able to model the corresponding table. The predictive performances, measured by the empirical risk, are given in Table 2.

Two facts could be considered: using a lemmatized vocabulary and removing the few frequent words had no impact on the predictive performance. When compared with the results of the first experiment, we moreover could observe that the co-clustering, by aggregating information, did not destroy the correlation between the threads and the forum categories.

Supervised modeling with augmented representation. – We led a third and last supervised experiment. We redone the previous experiment this time with considering an extra descriptive variable: the index of the group which the thread belonged to. For each vocabulary, we led a supervised modeling with `Khiops`. The predictive performances were measured by the empirical risk and reported to Table 3. Following the conclusion about the previous experiment, we limited ourselves to $k = 10$.

Vocabulary	k	number of attributes	empirical risk learning	empirical risk test
raw	2	25,789	34.67%	54.67%
	5	11,970	31.26%	54.27%
	10	7,249	33.75%	54.69%
lemmatized	2	20,889	31.47%	54.47%
	5	10,010	31.33%	54.53%
	10	6,198	30.97%	54.72%

Table 2: *Empirical risk for learning and test of the `Khiops` classifiers on a bag-of-words representation of the threads. The words whose frequency is above k were removed.*

Vocabulary	k	number of attributes	empirical risk learning	empirical risk test
raw	10	7,249	25.63%	51.43%
lemmatized	10	6,198	27.47%	50.17%

Table 3: *Empirical risk for learning and test of the `Khiops` classifiers on a bag-of-words representation of the threads, considering for each thread the index of its group.*

We observed that including the index of the group given by the co-clustering significantly (4.5%) improved the performance of the supervised model, for both raw and lemmatized vocabulary. The impact of lemmatizing stayed marginal (1.2%). We conclude that lemmatizing or filtering the less frequent words has a low impact: these steps can be avoided. Provided by a simple and non parametric method, the co-clustering is efficient and directly usable.

V. Conclusion

The animation of a forum is a heavy task with few automation. In this article, we aimed at exploring some ways of automation with the help of efficient and non parametric tools for data analysis. We proposed and applied an agnostic and automatic method for enhancing and evaluating the information in structured textual data.

The first conclusion is technical: firstly, lemmatizing the vocabulary or pruning it according to a minimum frequency threshold did not bring any supplementary information when associating a thread to a forum category. Secondly, the co-clustering algorithm correctly and comprehensively summarized the information contained in the first comment of each thread.

The second conclusion is functional: apart from the ability to automatically detect the spam, the clusters could enhance the use of the forum and its content, in order to increase the audience. For example, a keyword-oriented search in the content could lead to a theme-oriented search. A technical advisor could easily extract the content useful to build a tutorial.

The main appeal of the proposed methodology lies in the use of a co-clustering algorithm that avoids tedious text processing, while alternative approaches require a huge manual effort to extract the keywords.

References

[1] Arindam Banerjee, Srujana Merugu, Inderjit S. Dhillon, and Joydeep Ghosh. Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6:1705–1749, 2005.

[2] Mustapha Baziz, Mohand Boughanem, and Nathalie Aussenac-Gilles. The Use of Ontology for Semantic Representation of Documents. In Ying Ding, Keith van Rijsbergen, Iad Ounis, and Joemon Jose, editors, *The 2nd Semantic Web and Information Retrieval Workshop(SWIR), SIGIR 2004 , Sheffield UK, 29/07/04*, pages 38–45, juillet 2004.

[3] M. Boullé. Compression-based averaging of selective naive Bayes classifiers. *Journal of Machine Learning Research*, 8:1659–1685, 2007.

[4] M. Boullé. *Recherche d'une représentation des données efficace pour la fouille des grandes bases de données*. PhD thesis, Ecole Nationale Supérieure des Télécommunications, 2007.

[5] M. Boullé. Data grid models for preparation and modeling in supervised learning. In I. Guyon, G. Cawley, G. Dror, and A.R. Saffari, editors, *Hands-on pattern recognition*. Microtome, 2011.

[6] M. Boullé. Data grid models for preparation and modeling in supervised learning. In I. Guyon, G. Cawley, G. Dror, and A. Saffari, editors, *Hands-On Pattern Recognition: Challenges in Machine Learning, volume 1*, pages 99–130. Microtome Publishing, 2011.

[7] Laurent Candillier, Isabelle Tellier, Fabien Torre, and Olivier Bousquet. Cascade evaluation of clustering algorithms. In Johannes Furnkranz, Tobias Scheffer, and Myra Spiliopoulou, editors, *17th European Conference on Machine Learning (ECML'2006)*, volume LNAI 4212 of *LNCS*, pages 574–581, Berlin, Germany, september 2006. Springer Verlag.

[8] Inderjit S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *KDD '01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 269–274, New York, NY, USA, 2001. ACM.

[9] Inderjit S. Dhillon, Subramanyam Mallela, and Dharmendra S. Modha. Information-theoretic coclustering. In *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 89–98, New York, NY, USA, 2003. ACM.

[10] Bruno Guerraz, Marc Boullé, Dominique Gay, and Fabrice Clérot. Khiops coviz: A tool for visual exploratory analysis of k-coclustering results. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2014, Nancy, France, September 15-19, 2014. Proceedings, Part III*, pages 444–447, 2014.

[11] L. R. Khan. *Ontology-based Information Selection*. PhD thesis, University of Southern California, 2000.

[12] Noam Slonim and Naftali Tishby. Document clustering using word clusters via the information bottleneck method. In *SIGIR '00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 208–215, New York, NY, USA, 2000. ACM.