Discontinuity, Nonlinearity, and Complexity 1(1) (2012) 1-6

Discontinuity, Nonlinearity, and Complexity

https://lhscientificpublishing.com/Journals/DNC-Default.aspx

Q-analysis based clustering of online news

David M.S. Rodrigues
david.rodrigues@open.ac.uk

Centre for Complexity and Design, Faculty of Mathematics, Computing and Technology,
The Open Univeristy, Milton Keynes, MK7 6AA, UK

Submission Info

Communicated by Referees
Received DAY MON YEAR
Accepted DAY MON YEAR
Available online DAY MON YEAR

Keywords

Q-analysis
Community detection
Online News
Topic Modelling

Abstract

With online publication and social media taking the main role in dissemination of news, and with the decline of traditional printed media, it has become necessary to devise ways to automatically extract meaningful information from the plethora of sources available and to make that information readily available to interested parties. In this paper we present a method of automated analysis of the underlying structure of online newspapers based on Q-analysis and modularity optimisation. We show how the combination of the two strategies allows for the identification of well defined news clusters that are free of noise (unrelated stories) and provide automated clustering of information on trending topics on news published online.

1 Introduction

Every day the Internet presents a huge amount of new information, either in personal or institutional web pages, posted in social networks or as a reflex of the media (TV, newspapers) regular diffusion of news. Regarding this last case, information is published as a result of complex relations between facts and news, shaping a ever-changing network of relations between topics. Topic detection and the unveiling of dynamic relations between topics can give new insights to the understanding of communication mechanisms in human societies.

In recent years there has been a surge of interest in term extraction and automated text categorisation [1–3], mainly because of the increasing amount of information that is being produced online. Examples include extracting and classifying biological text [4], the categorising of online news [5] and personalised recommendation systems [6]. With online information's rapid growth also has come the development of automated and agile methods to process it. The literature provides many examples of term extraction methods that can be categorised roughly into two fields, according to their different perspectives. One approaches the task from a linguistic, terminology and natural language processing perspectives [7], and the other uses mainly tools from the statistical and information retrieval fields [2, 8].

ISSN 2164 – 6376, eISSN 2164 – 6414/$-see front materials © 2012 L&H Scientific Publishing, LLC. All rights reserved.
DOI : 10.5890/DNC.2012.02.001

On the other hand, complex systems are composed of n-ary relations [9] and should be described in languages that support these high dimensional relations. By including structural information on the connectivity of the graph representing the problem one can filter out those nodes that otherwise would be misclassified by traditional algorithms.

In this work we use a hybrid connectivity based approach that combines Q-analysis [10–14] and Modularity based clustering [15, 16] to obtain an automated procedure for clustering news based on the connectivity of their structure.

This paper is organised in five sections. After this introductory section some of the most important related work is presented in section 2. Then, section 3 shows the dataset used for analysis and explains the algorithm developed for the hybrid connectivity based approach shown in this paper. In section 4 the results of the analysis are presented and in section 5 an overview and implications of this work is carried out.

2 Related work

Recent research in the domain of topic detection applied different techniques, including regression models, nearest neighbor classification, Bayesian probabilistic approaches, decision trees, inductive rule learning, neural networks, online learning and Support Vector Machines [1, 3, 17–20]. However, most of those approaches are supervised and require a training set, where documents previously classified by humans are used as input to make the system learn each category's particular features. These approaches face two major restrictions: they are language-dependent, requiring the work of specialists for analysing and classifying; and they are not adapted for finding new categories in data without re-training.

Other attempts to solve the above problem include using a dynamic network with a time sliding window where changes between consecutive generated networks are evaluated as the variation of information between consecutive windows [21, 22].

Also, previous community detection algorithms in graphs aimed at the partition of all nodes of the graph into disjoint clusters. This process was traditionally considered exclusive so that a node belonging to one cluster wouldn't belong to another. To overcome this restriction several algorithms have been proposed that allows overlapping of communities. One example of such algorithms is the clique percolation method by [23, 24]. On the other end of the spectrum, very little attention has been given to to the cases where not all nodes should be included in a cluster. These nodes nodes aren't really associated with any other node in a meaningful way and traditional clustering algorithms don't have mechanisms to deal with them. The main reason for this is that they lack the structural information needed to include or reject them from the clustering process. Each edge is unidimensional representing a binary relation and the only additions to this simplification are the inclusion of directionality and weights to these relations.

Q-analysis[a] is a mathematical framework to formalise the structure of a relation between sets. It was developed and introduced to the social sciences by English mathematician Ronald Atkin and colleagues in the early 1970s and has been used as a research methodology in a diverse range of areas and contexts. Most of Q-analysis ideas are found in the initial reports and papers Atkin wrote during the development of the project "The Urban Structure Research Project" 1971-1974, in which he wanted to answer a set of questions [10, 11, 25, 26]:

- How does the structure come into being?

[a]The term Q-analysis first appears in Atkin's theoretical paper "From cohomology in physics to q-connectivity in social science" [10, p.156]

- What are the components of the structure?
- How does a structure change with time?

In this work we propose a novel procedure for the automatic classification of newspaper articles based on the bipartite graph generated from the newspaper articles published online and from the tags associated with them entered by journalists at publication time. We analyse this system by first using Q-analysis on the graph of news and then by clustering the higher q-connectivity induced subgraphs. After extraction of this high connectivity graph one can proceed to cluster the news stories by using any traditional graph clustering algorithm [27, for in-depth review]. In this case we show the results of applying a modularity optimisation [15] based method to the analysis of the news published online by *The Guardian*[b] and how the Q-analysis improves the quality of the clustering.

3 Materials and Methods

The Guardian classifies every news item published with a set of metadata that can be used for clustering information. The two most interesting metadata fields are the **section** and the **tag** metadata. Each document has one **section** field corresponding to the section of the newspaper where the story was published and one or several **tag** fields that the journalist or editor chose to characterise that particular story. The algorithm presented here takes advantage of this human labelling in characterising the structure of the resulting network created by all the news from *The Guardian*.

Guardian articles and tags were collected each month in 2010, 2011 and 2012. From these thirty six data sets were created of the monthly article-tag bipartite relation. Typically there were about 1,500 articles per month with 2,000 to 4,000 tags.

For any particular month let A be the set of articles and T be the set of tags. Let $\sigma(a_i) = \langle t_0, t_1, \cdots, t_p \rangle$ be defined to be a p-simplex for the article-tag relation, where a_i is related to t_j for $j = 0, 1, \cdots, p$. The tags t_j are called vertices. For example, σ("Iraq bombings kill dozens") = ⟨Iraq, Middle East and North Africa, World news, Al-Qaida⟩ is an example of an abstract simplex with four vertices. In general a p-simplex has a geometrical realisation as a polyhedron in multidimensional space so that the simplex σ("Iraq bombings kill dozens") can be represented by a tetrahedron with four vertices in three-dimensional space. In general the dimension of a simplex is the number of its vertices plus one. For example the dimension of a line (2 vertices) is 1, the dimension of a triangle (3 vertices) 2, the dimension of a tetrahedron (4 vertices) is 3, and so on. High dimensional articles have many tags as vertices and low dimensional articles have few tags. In Atkin's theory, two simplices are q-near when they share a q-dimensional face with $q+1$ vertices. A set of simplices with all its faces is called a simplicial complex. Two simplices are q-connected if there is a chain of pairwise q-near simplices between them. Being q-connected is an equivalence relation on a set of simplices and partitions them into q-connected components. A list of the q-components for a set of simplices is called Q-analysis. For more details see [9, 28].

Networks, or q-graphs [9] can be constructed from the bipartite relations by letting each simplex be a vertex, and connecting a link between vertices if they are p-near, for $p \geq q$. The resulting networks were studied in terms of nodes count, number of clusters, modularity of the resulting clustering, maximum cluster size, and fraction of vertices present in the resulting graph.

For the purpose of using a community finding algorithm a fast greedy algorithm proposed by [15] was chosen. This algorithm is a hierarchical agglomerative algorithm for detection of community structure that

[b]The Guardian (`http://www.guardian.co.uk/`)

is very fast. The algorithms runs in $O(md\log n)$ for a network with n vertices, m edges and where d is the depth of the dendrogram. For networks that are hierarchical this means that $d \sim \log n$ and if the networks are also sparse, then $n \sim m$ making the running time essentially linear at $O(n\log^2 n)$. This allows a quick overview of the community properties present in the graph of tags[c].

The fast greedy algorithm is an agglomerative algorithm and by plotting the modularity of the communities found at each merge one can find at which iteration of the merging process the maximum value of modularity is obtained. This is the point at which one selects the cut in the dendrogram and identifies the communities present.

Algorithm 1 Pseudocode for the automated news clustering and filtering algorithm

Require: *RSS feed*
  /*Initialization*/
  for all *news item* $\in$ *RSS feed* **do**
    Fetch HTML page
    Convert to Text
    Extract Tags
    Add news story to bipartite graph
  end for
  for all values of q in bipartite graph **do**
    Extract induced graph G_i with simplexes that are at least q-near with other simplex $Q \geq q$
    Use modularity optimisation to compute modules in graph G_i
  end for
  Rank graphs (G_i) according to modularity index obtained from clustering algorithms

A pseudocode for the algorithm of automatically processing (filtering and clustering) the monthly news published by *The Guardian* is outlined in Algorithm 1. After retrieval of the news story, usually from the RSS feed, the html file is then converted to text and the tags associated with that particular news are retrieved. Each story is added to the bipartite graph and after this step the resulting bipartite graph between news stories and tags of *The Guardian* is analysed by extracting the induced subgraphs of stories that were at least q-near to other news. For this we processed the q-graphs of the news-tags bipartite graph for values of $q > 0$.

4 Results

Initially a graph is constructed by defining each published story as a node in the graph and then by considering that two stories were connected if they shared at least one tag among them. This approach can be thought of as the construction of the 0-graph from the simplicial complex of the bipartite article-tag relation. Articles are vertices and there is a link between them if they share at least 1 tag (are 0-near). This corresponds to the simplest projection with the broadest structural information. As an example, the application of community detection algorithms based on modularity optimisation to the graph of one month generates a clustering where the maximal value of modularity is 0.48 for a total of 9 communities as shown in table 1.

[c]The processing is done in R [29] using the package *igraph* [30] that has implemented many useful algorithms for graph manipulation.

Table 1 Cluster sizes for *The Guardian* news published during the month of November, 2011

id:	1	2	3	4	5	6	7	8	9
size:	363	303	96	221	6	5	102	13	46

This clustering reveals large components with more then 300 news items (clusters 1 and 2), meaning that probably a division into more components would be of greater interest and that these large clusters do not capture the fine structure of the news, due to the resolution limit of modularity optimisation methods [31]. For this we need a new approach, one that effectively can give insights into the news structure and that removes *noise news*[d] that otherwise would be part in spurious classifications.

The previous example analysis of the tags network of *The Guardian* news shows clearly that calculating the clusters via modularity optimisation alone isn't enough. In it we considered full connectivity between nodes of the graph (two nodes where connected if they shared at lease one tag). We now show that by extracting the induced subgraph of higher connected nodes we can obtain a clearer separation of modules in the news structure.

For this we proceeded with a Q-analysis of the news-tags bipartite graph. In Q-analysis [28, 32, 33] two nodes are connected by a link if they share at least $(q+1)$ common attributes. In this case two news are connected if they share at least $q+1$ tags. The resulting networks constructed in this way guarantee a minimum level of connectedness between published news. From the resulting structure the induced graphs are extracted by thresholding the minimum connectedness between news.

Each of our thirty six bipartite article-tag relations gives a set of simplicies, $\{\sigma(a_i)|a_i \text{is an article for this month}\}$. The proportions of these simplicies of different dimensions is shown on a Cartesian graph in Figure 1 (in fact all thirty six Cartesian graphs are shown superimposed in Figure 1. This shows that the majority of articles (as simplicies) have low dimensions of about $q = 5$ (6 vertices) or less.

A characterisation of the induced subgraphs in terms of several properties (following figures) as a function of Q helps understand the structure and to filter the noise news that create spurious relations among the dataset.

The first aspect to take into consideration when applying this method is the number of vertices (news articles) that is filtered when considering the q connectivity of the system. One can put this in other terms by considering the fraction of vertices that remain after filtering out all that do not have at least dimension q. In Figure 1, 80% of news are still connected for $q = 3$. As the value of q increases, the fraction of retained news drops quickly. For a connectivity of $q = 10$ the resulting structures retain less than 20% of the original nodes. This sharp drop of the fraction of retained nodes is indicative that the majority of the news items are connected through low dimensional faces ($q = 3, 4, 5$).

Figure 2 shows statistics for the q-graph. There is one such graph for each of the 36 study months. In these q-graphs the vertices are those articles related to q or more tags, and there is a link between two articles if they share q or more tags. The q-graphs all have a giant component. Figure 2 shows the proportion of articles in the giant component by dimension q.

In Figure 2 it is observed that for low values of connectivity there is a giant component that captures most of the nodes in the structure, but for values of $q >= 5$ the fraction of nodes present in the maximal cluster drops substantially and the giant component disappears. This transition from giant component to the more regular components is also indicative of a drastic change in the connectivity of the resulting structures

[d] news that aren't highly connected to others and therefore can be discarded when considering the task of understanding the main themes of a newspaper

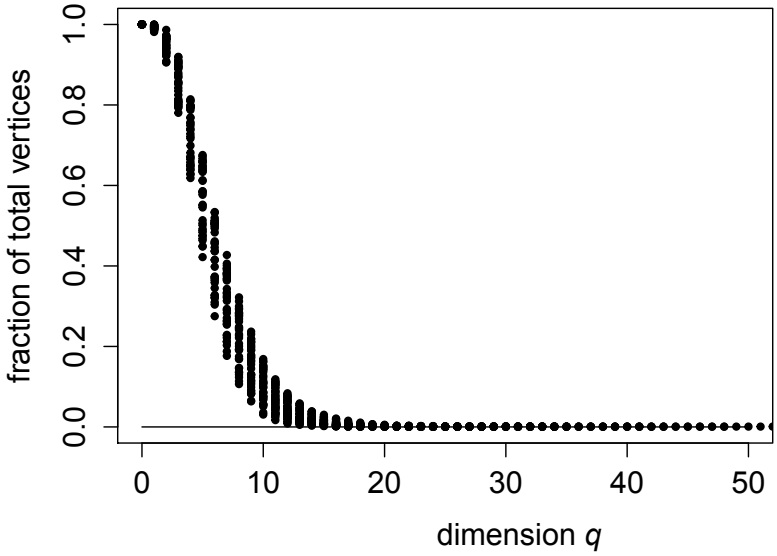

Fig. 1 Fraction of vertices in the resulting q-graphs as a function of q for each of the 36 months (each dot in the figure represents one month)

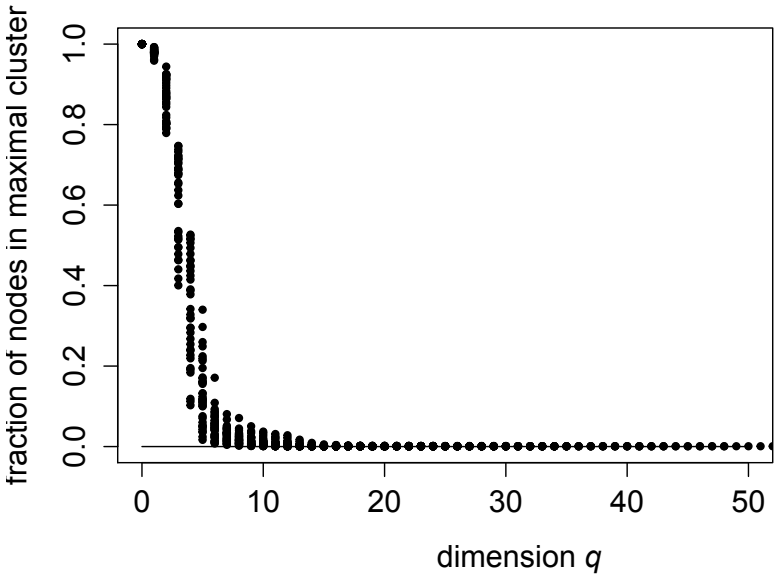

Fig. 2 Fraction of nodes in the maximal cluster as a function of q for each of the 36 months (each dot in the figure represents one month)

after filtering the structure of news of low connectivity news. Naturally this is expected as q increases the percolation threshold is met and the bipartite graph is decomposed into multiple small units.

The low value at which this transition is observed provides evidence for the low connectivity of different stories and might be a reflection of the way the tagging mechanism works. Maybe journalists tend to give each story a few specific keywords instead of more general ones that would be broader in meaning (although increasing the connectivity of the news story).

For the purpose of filtering content that is not highly connected and therefore unrelated to other news it

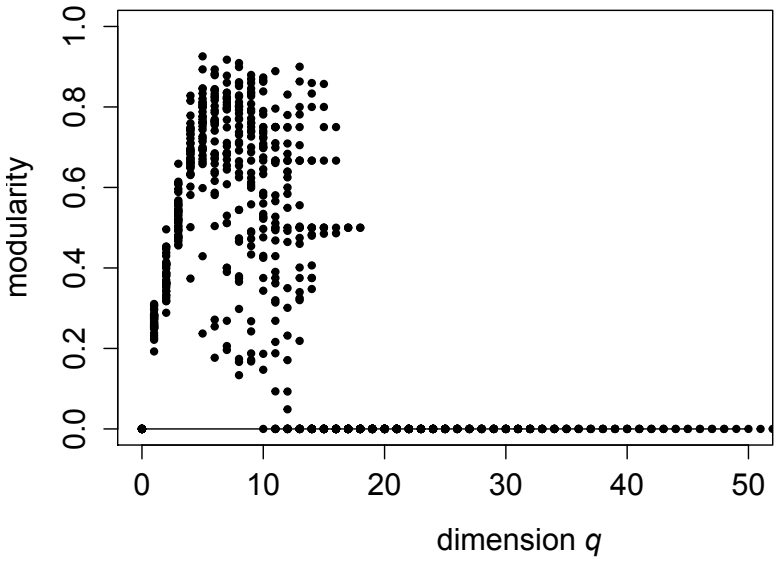

Fig. 3 Modularity of the induced graph as a function of q for each of the 36 months (each dot in the figure represents one month)

is important to choose a value of q as a cutoff to optimise some properties of the resulting induced graph. These properties include the retention of a majority of nodes, high number of clusters, relatively small size of the maximal cluster and high modularity from the resulting graph. Figure 3 shows that after processing each induced graph with the agglomerative hierarchical algorithm by Clauset *et al*. [15], that the resulting induced graph presents high values of modularity for $q \geq 3$. According to Newman, a network that presents modularity above 0.3 is considered to have some sort of community structure, meaning that the density of edges intra communities is much higher than that the density of the edges inter communities (obviously, when compared against a null model where a network is constructed randomly with the same number of edges/nodes and degree distribution).

It is clear that to identify the most significant components in the news published one is interested in some useful characteristics of the induced graph. It should have high clustering, high number of components, and also high degree assortativity. It is expected also that the value o modularity for this graph is also high. Usually one wants to choose an induced graph such that the number of nodes present is still high. In the case of *The Guardian* choosing $q = 5$ presents these characteristics.

Figure 4 shows the calculated number of clusters using the modularity optimisation algorithm as a function of q for each month of the three years os news. It is observed in Figure 4 that the number of communities quickly rises up to $q = 5$ and then drops again. This is due both to the increase in modularity of the resulting induced subgraph and to the nearness of the percolation threshold where the giant component disappears and many small components are present.

This can be observed in an example of the month of November 2011 in Figure 5 where you the communities identified for the induced subgraph ($q = 5$) and it is visible the presence o many isolated modules with 2 or 3 news items.

When comparing the distribution of the size of the modules of this November month with that of table 1 we can see that we have now many more fine detailed clusters. Figure 6 shows the size distribution of the clusters in the induced subgraph of November 2011 news from *The Guardian*.

The clustering process gives higher values for modularity of the resulting induced graphs as shown in

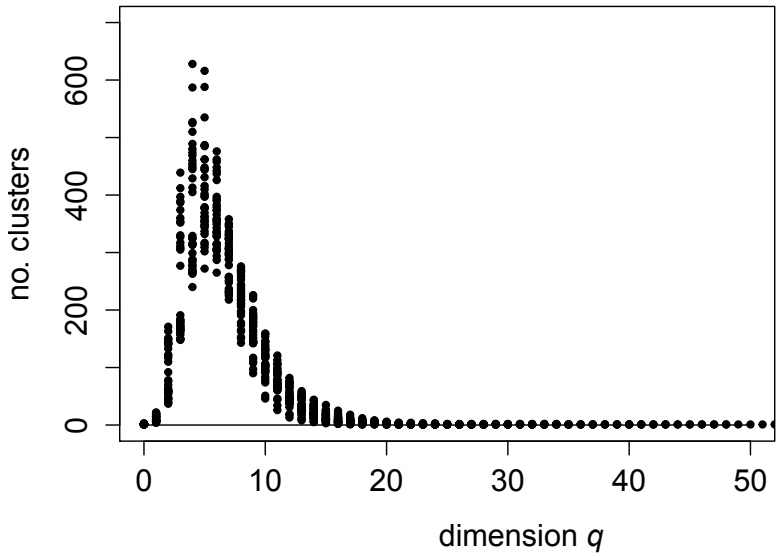

Fig. 4 No. of clusters as a function of q for each of the 36 months (each dot in the figure represents one month)

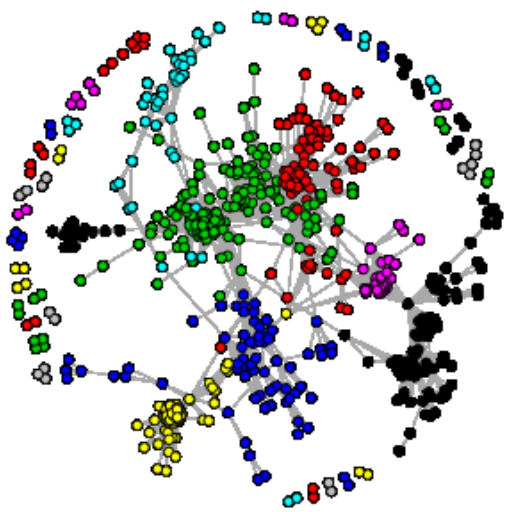

Fig. 5 *The Guardian* news, coloured by community: $q = 5$, $Modularity = 0.71$

Figure 3. This gives high confidence on the structural properties of the clusters identified. Manual inspection of the clusters was conducted and revealed a total absence of unrelated stories in the identified clusters. By using this system the we were able to identify 48 clusters instead of the previous 9 with higher modularity.

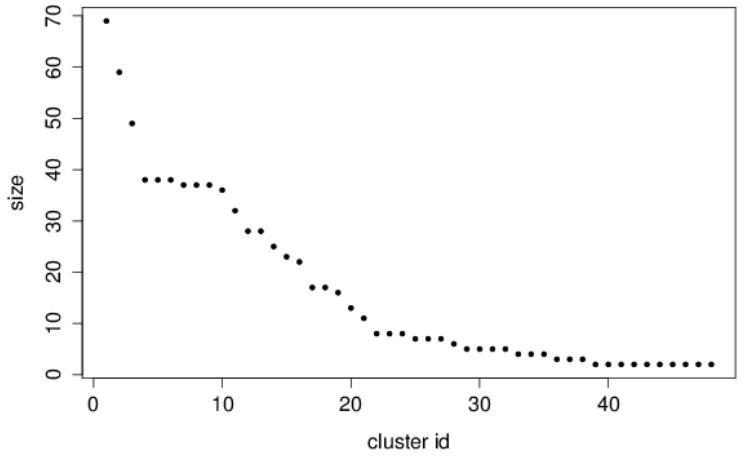

Fig. 6 Size distribution of clusters of *The Guardian* news ($q = 5$) for the month of November 2011

This is an indication that the previous modularity optimisation had limitations (mainly because of the noise stories) and it was merging into the same cluster unrelated news stories. The size distribution shows the presence of many small sized cluster (typically between 2 and 5 news) but at least 10 clusters are constituted by 30 or more news stories. These are the main topics being discussed during that month and include the scandal of James Murdoch news empire, the evolution of the Syrian situation and the Arab Spring and the Eurozone debt crisis, among others.

The observed difference in modularity is possible because low connected stories act as bridges between different topics via general tags as *world news*, or *sport*, that are not specific enough and hinder the identification of correct modules in the news. By removing those stories with Q-analysis, the quality of the resulting community detection is substantially improved as the algorithm don't need to classify stories that are naturally unrelated to the topics of of the news of that period.

5 Discussion

In this paper we present a novel way for filtering news stories published in online newspapers by using Q-analysis and modularity optimisation. The process deals with the problem of selecting which relevant stories should be considered when clustering documents. We show how a bipartite graph formed from the documents and the tags associated with each document can be used to filter the stories that aren't strongly connected to other members of the graph.

The main advantage of using this technique is that the process of analysis is automated. On the other hand the process poses the disadvantage of discarding those nodes that aren't highly connected, but this disadvantage is a parameterised one and can be minimised by lowering the value of connectedness needed for inclusion in the induced subgraph. If all nodes are to be included then the problem reverts to a traditional clustering problem.

This method presents itself to many applications where the structural properties of the system are of high importance. Traditional clustering algorithms traditional consider the clustering process on graphs where nodes are equivalent and where the different structural connectivities of the nodes isn't taken into

account. By using extra information from the bipartite graph, through Q-analysis, it is possible to filter relevant information from spurious information in an automated way.

In the case of the corpus presented here, it relies on human labelling of the documents, but the construction of the bipartite graph can be done with one of the many automated topic modelling strategies developed in recent years [34, 35].

Another advantage of this method is that it can be applied to different corpora of textual based documents. Instead of analysing a single newspaper, it can be applied to daily news from different newspapers or channels to extract relevant stories and topics. It provides a way of clustering relevant trends in the news almost in real time. The method can also be applied to other fields: automated clustering of new scientific publishing, tracking of opinion dynamics on social media, and detection of brand awareness in online communities, constitute potential applications where knowledge of the structural properties of the system improves the quality of the analysis.

References

[1] Y. Yang and X. Liu, "A re-examination of text categorization methods," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '99, pp. 42–49, ACM, 1999.

[2] P. Pantel and D. Lin, "A statistical corpus-based term extractor," in *Biennial Conference of the Canadian Society on Computational Studies of Intelligence: Advances in Artificial Intelligence*, pp. 1–10, Springer-Verlag, 2001.

[3] A. Cardoso-Cachopo and A. L. Oliveira, "An empirical comparison of text categorization methods," in *String Processing and Information Retrieval* (M. A. Nascimento, E. S. D. Moura, and A. L. Oliveira, eds.), pp. 183–196, Springer Verlag, Heidelberg, DE, 2003.

[4] M. Lee, W. Wang, and H. Yu, "Exploring supervised and unsupervised methods to detect topics in biomedical text," *BMC bioinformatics*, vol. 7, pp. 140–140, 2006.

[5] T. Weninger and W. Hsu, "Text extraction from the web via text-to-tag ratio," in *19th International Workshop on Database and Expert Systems Application, 2008. DEXA'08.*, pp. 23–28, IEEE, 2008.

[6] Z.-k. Zhang, T. Zhou, and Y.-c. Zhang, "Personalized recommendation via integrated diffusion on user-item-tag tripartite graphs," *Physica A: Statistical Mechanics and its Applications*, vol. 389, pp. 179–186, 1 2010.

[7] V. Hatzivassiloglou, L. Gravano, and A. Maganti, "An investigation of linguistic features and clustering algorithms for topical document clustering," in *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, (Athens, Greece), pp. 224–231, 2000.

[8] K. Nigam, J. Lafferty, and A. McCallum, "Using maximum entropy for text classification," in *IJCAI-99 Workshop on Machine Learning for Information Filtering*, pp. 61–67, 1999.

[9] J. H. Johnson, *Hypernetworks in the science of complex systems*. Imperial College Press (London), 2013.

[10] R. H. Atkin, "From cohomology in physics to q-connectivity in social science," *International Journal of Man-Machine Studies*, vol. 4, no. 2, pp. 139 – 167, 1972.

[11] R. H. Atkin, *Mathematical Structure in Human Affairs*. 48 Charles Street, London: Heinemann Educational Publishers, 1 ed., 1974.

[12] J. R. Beaumont and A. C. Gatrell, *An introduction to Q-analysis*. Norwich Norfolk: Geo Abstracts, 1982.

[13] J. H. Johnson, "A survey of q-analysis, part 1: The past and present," in *Proceedings of the Seminar on Q-analysis and the Social Sciences, Universty of Leeds*, 9 1983.

[14] D. M. S. Rodrigues, "Identifying news clusters using q-analysis and modularity," in *Proceedings of the European Conference on Complex Systems 2013* (A. Diaz-Guilera, A. Arenas, and Á. Corral, eds.), (Barcelona), 9 2013.

[15] A. Clauset, M. Newman, and C. Moore, "Finding community structure in very large networks," *Phys. Rev. E*, vol. 70, p. 066111, Dec 2004. Phys. Rev. E 70, 066111 (2004).

[16] M. Newman, "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences*, vol. 103, no. 23, pp. 8577–8582, 2006.

[17] Y. Miao and X. Qiu, "Hierarchical centroid-based classifier for large scale text classification," in *Large Scale Hierarchical Text classification (LSHTC) Pascal Challenge*, 2009.

[18] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," *Machine*

Learning ECML98, vol. 1398, no. 23, pp. 137–142, 1998.

[19] M. Hamamoto, H. Kitagawa, J. Pan, and C. Faloutsos, "A comparative study of feature Vector-Based topic detection schemes a comparative study of feature Vector-Based topic detection schemes," in *Web Information Retrieval and Integration, 2005. WIRI '05. Proceedings. International Workshop on Challenges in*, pp. 122–127, IEEE, Apr. 2005.

[20] R. V. Solé, B. Corominas-Murtra, S. Valverde, and L. Steels, "Language networks: Their structure, function, and evolution," *Complexity*, vol. 15, no. 6, pp. 20–26, 2010.

[21] M. Meilă, "Comparing clusterings—an information based distance," *J. Multivar. Anal.*, vol. 98, no. 5, pp. 873–895, 2007.

[22] D. M. S. Rodrigues, "The observatorium – the structure of news: topic monitoring in online media with mutual information," in *Proceedings of the European Conference on Complex Systems* (J. Louçã, ed.), (Lisbon), Complex Systems Society, 9 2010.

[23] I. Derenyi, G. Palla, and T. Vicsek, "Clique percolation in random networks," *Physical Review Letters*, vol. 94, p. 160202, 2005.

[24] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature*, vol. 435, pp. 814–8, 6 2005. PMID: 15944704.

[25] R. H. Atkin, R. Bray, and I. Cook, "A mathematical approach towards a social science," *Essex University Review*, no. 2, pp. 6–8, 1968.

[26] R. H. Atkin, J. Johnson, and V. Mancini, "An analysis of urban structure using concepts of algebraic topology," *Urban Studies*, vol. 8, no. 3, pp. 221–242, 1971.

[27] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3-5, pp. 75–174, 2010.

[28] J. H. Johnson, "Some structures and notation of Q-analysis," *Environment And Planning B*, vol. 8, pp. 73–86, 1981.

[29] R Development Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0.

[30] G. Csardi and T. Nepusz, "The igraph software package for complex network research," *InterJournal*, vol. Complex Systems, p. 1695, 2006.

[31] S. Fortunato and M. Barthelemy, "Resolution limit in community detection," *physics/0607100*, 7 2006. Proc. Natl. Acad. Sci. USA 104 (1), 36-41 (2007).

[32] J. H. Johnson, "Multidimensional multilevel networks in the science of the design of complex systems," in *ECCS 2005 Satellite Workshop: Embracing Complexity in Design* (J. Johnson, ed.), vol. ECCS 2005 Satellite Workshop: Embracing Complexity in Design, 2005.

[33] J. H. Johnson, "Can complexity help us better understand risk?," *Risk Managment*, vol. 8, no. 4, pp. 227–267, 2006.

[34] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[35] W. Li and A. McCallum, "Pachinko allocation: Dag-structured mixture models of topic correlations," in *Proceedings of the 23rd international conference on Machine learning*, ICML '06, (New York, NY, USA), pp. 577–584, ACM, 2006.