European Journal of Molecular & Clinical Medicine
ISSN 2515-8260 Volume 07, Issue 02, 2020

INTELLIGENT GENETIC HYBRID CLUSTERING DATA ANALYTICS METHOD IMPLEMENTING EFFICIENT NEWS FEED SERVICES IN ONLINE MOBILE APPLICATIONS

[1]M P Rajakumar, [2] B.Ramathilagam, [3]Shanthi D L, [4]R.Senthil, [5]Dr.Syed Khasim

[1]Associate Professor, Computer Science and Engineering, St.Joseph's College of Engineering, OMR, Chennai - 600 119, Tamilnadu.

[2]Associate Professor, Electronics and Communications Engineering, Mailam Engineering College Mailam Tindivanam (Tk), Villupuram District.604304.

3Assistant professor, Information Science and Engineering, BMS Institute of Technology and Management, Avalahalli, Yelahanka, Bangalore-64.

[4]Assistant Professor, Information Technology, SRM Institute of Science and Technology, Delhi-NCR Campus, Delhi-Meerut Road, Modinagar, Ghaziabad-201204, Uttar Pradesh.

[5]Professor, Computer Science & Engineering, Dr.Samuel George Institute of Engineering and Technology, Markapur, Prakasam DT-523316, India.

Abstract:

This demonstration concerns a system designed and implemented to automatically build multimodal aggregations of informative news items coming from the two domains of digital television and the web. Though in recent times several technological solutions have addressed the problem of clustering online articles, little is available which is capable of integrating these two sources of information. The demonstrated system is based on a novel hybrid clustering approach able to construct a directed graph of items representing a certain event (e.g., the tour of the Olympic torch), and to deliver an RSS service in which each item includes contributions coming from the two domains. In-house experimental evaluations have already proved the efficacy and accuracy of the service.

Index: Intelligent Genetic Model, Multimodal Aggregation, Hybrid Clustering, RSS Service, Data Analytics

I. Introduction

The availability of large multimedia collections now available for free is an important opportunity for the general public. However, access to this information, which is increasing day by day, is becoming a critical issue because the inherent diversity of data resources based on delivery media undermines processes aimed at extracting meaningful information from them. An example of this can be seen in the provisional publication of news related to current events, such as television and the web, especially in the informative news domain. The ability to auto-aggregate on these two domains refers to the novel-tie presented by our system in the current technical context. The functional architecture of the system is shown in Figure 1.

The system is a processing machine consisting of two input stream processing networks and an output processing network. The two input stream processing networks operate on digitalized television streams (DTV) and online newspaper feeds (RSSF), respectively, and produce the chain output chain Multimodal Aggregation Service (MMAS). The DTV stream was first analyzed and divided into programs using the automatic video clip detection technique described in Section 2.2. It then automatically divides into primary news items and applies different indexing methods to the entire program and individual items as described in Section 2.3.

Therefore, the final product of the DTV Stream Processing Network is Index Structure TV, which is an automatically recognized and processed television news item. More details on the implementation of the DTV Stream Processing Network [2]. The RSSF stream is generated through the RSS feeds of many popular online newspapers. Language parsing is done on each RSS topic with the aim of identifying the main elements (i.e., titles and descriptions) of the RSS descriptive elements from a language perspective. Conducts an investigation and initiates the resulting query into the DTV chain pro-ducked index structure. The details of this process are given in Section 2.5.

The hybrid clustering process is based on the results of high-automatic built-in queries following the technique described in Section 2.5. The resulting compilations are stored in MAI, which is a properly indexed data structure that acts as a data persistent collection from which the final MMAS is generated and sent to the users. To represent each cross-space aggregation, we have chosen to use the title of the online article related to the aggregation representative component according to the specification given in Section 3 of our system.

2. Programme Boundary Detection

We use optical-size video clip matching technology to automatically split live streams into programs. Video elements (shots) indicating the beginning and end of programs are used as reference prototypes for searching through captured video streams. Technology consists of a learning phase and an innovation phase. During the study phase, E, e.g., for each gross event of interest. Jingle Launched Program, NE Examples {e1, e2,. . . , ENE selected from daily television program acquisitions. In each case an adaptive threshold shot detection algorithm based on the ei displaced frame luminance difference (L-DFD) is implemented.

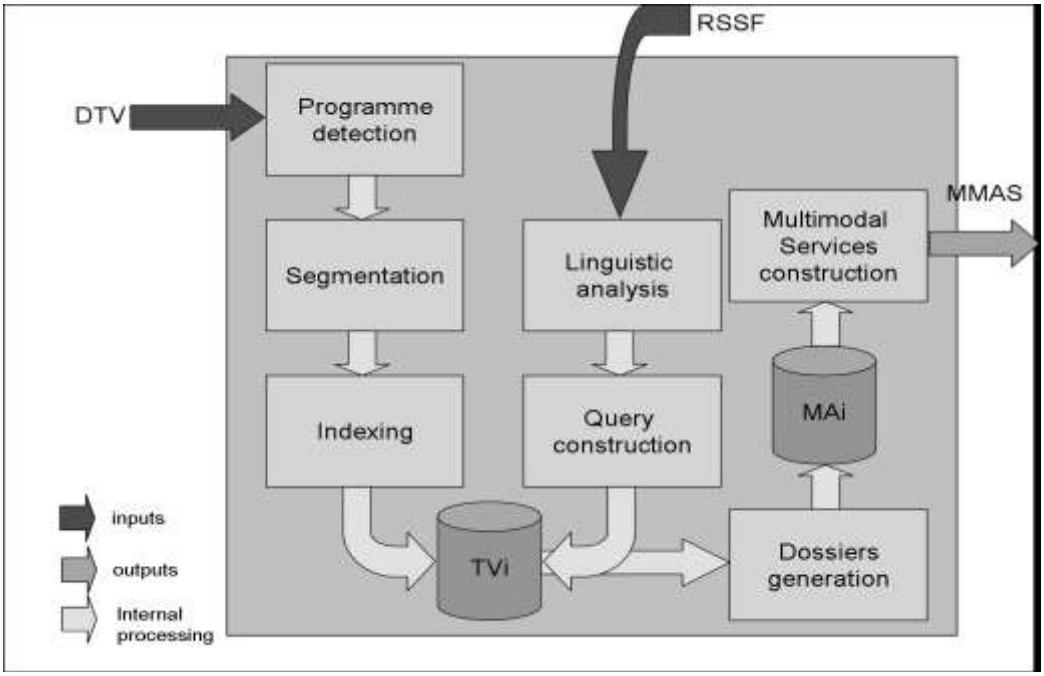

Figure 1.Intelligent Genetic Hybrid Cluster architecture

Adaptively is based on the local statistics of L-DFD, so that content having higher L-DFD variance is processed against higher thresholds. Let F = {f1, f2, . . . , fNf } be the set of low level visual features extracted from each shot. Each element of F is represented using a value space made up of normalized and uniformly distributed B-bin histograms, in which the last bin is used to count the pixels of the image for which the measurement of feature returns an undetermined value (e.g., hue for grey pixels). Therefore, for each macro-event E, we have a set AE of NE arrays of feature vectors of size Ns, AE = {a1, a2, . . . , aNE }. Each element am, n = 1. . . Ns of each array aAE is a real three-dimensional matrix of dimensionality Nf × B × ln. We compute two additional arrays aNE+1 and aNE+2, to store the mean and the standard deviation of extracted features [5].

To improve detection capability, we make a selection of the most promising NS shots associated with the macro-event. The choice is based on measuring the distance between the distribution of feature vectors collected from each NS shot and the normal distribution of the same properties calculated in the sample group of random shots. At the detection stage, the average feature vectors are used as static references, comparing the shots obtained using the distance measurement based on the histogram inter-section.

3. Implementation – Hybrid Modeling

Following the procedure described in the previous section, Newscast programs are divided into their logical sections, i.e., news articles, which are the basic elements stored in the TV index (see Figure 1). Dividing neuralcast programs into news articles using verbal and visual cues with the help of three layered heuristic frameworks. The heuristics used are based on an overview of the stylistic language of important programs (≈ 80 programs) taken at limited intervals from the daily schedules of the 7 major national broadcast channels. Basic Heuristics (H1) treats the boundaries of shots with anchorman as equivalent to news articles.We use another heuristics (H2) to find the anchorman shots, the speaker being the highest commentator and he speaks many times during the program and the time limits distributed by the program timeline.

This supervised speaker clustering process results in the speaker labeling all the speakers in the program and associating them with temporary sections of content. However, without interruption to external works, the application of the first two heuristics is not yet sufficient to identify the contexts in which Anchormann presents several short stories in a row. We use the third heuristic (H3) to overcome this limitation, i.e. we know that in most cases the introduction of a new short story involves a change in the camera shot (e.g., from a close-up shot to a wide one). Tracking this camera change gives the final clue to the partition.

Therefore, we perform a video shot clustering process to ensure the accuracy of the partition. This allows us to identify and classify shot clusters for studio shots containing Ank-hormone according to the same frequency / extension heuristic used to identify the candidate speaker (H2). This dual clustering process (both audio and video) allows a very simple and effective algorithm to select video and audio clusters based on mutual coverage percentages. Once detected, news items are represented using automatic speech recognition based on data models.

To achieve automatic segmentation of live streams into programs we make use of an optimized video clip matching technique. Video elements (shots) indicating starting and ending of programs are used as reference prototypes to be searched through the acquired video streams. The technique includes a learning phase and a detection phase. In the learning phase, for each macro-event of interest E, e.g. a programme starting jingle, NE instances {e1, e2, . . . , eNE } are selected

from daily television program acquisitions. On each instance ei an adaptive threshold shot detection algorithm based on displaced frame luminance difference (L-DFD) is performed.

A. Automatic Query Construction

In our system RSS information items are elaborated by a part of speech tagger based on [3] which, in opposition to statistical approaches, extracts elements of the title and description sentences which are important from the linguistic point of view. The extracted elements are then used to build full text queries which are run on the index provided by Lucene 1 on the speech transcripts of the television news items.

B. The Core Aggregation Mechanism

To address the task of multimodal aggregation, we propose a hybrid clustering algorithm based on the principle of semantic relevance and on a vector projection n similarity measurement between information items.

Definition 1 An information item I1 is semantically relevant to another information item I2 if the consumption of I1 by a consumer c satisfies the expectations of c about I2.

Let C1 = {∀iI1i} and C2 = {∀jI2j} be two sets of information items with cardinalities |C1| and |C2| respectively, for which a distance metric in the space C = C1 ∪C2 is not defined.

The key idea of our method lies in the definition of an vector projection similarity measurement S in the space C1 (and similarly in the space C2) as follows:

$$S : C1 \times C1 \rightarrow [0,\infty) \quad (1)$$

$$S(a, b) = VR(a) \cdot VR(b) - VR(a) \quad (2)$$

$$VR(x) = (R(x, y1), R(x, y2), \ldots, R(x, y|C2|)) \quad (3)$$

Where C1,C2 is the norm induced by the inner product in the VR() space. Elements of vector VR(x) are the ordered set of values of a semantic relevance function R : C1 → C2 (see Section 2.5.1) of the information element x w.r.t. all elements of C2. Given the intrinsic asymmetrical nature of Equation 2, we can construct graphs as that depicted in Figure 2 by imposing a threshold α on the values of S(a, b). Figure 2 illustrates an example of an aggregation, taken from the ones actually detected by our system. In the Figure, arrows between two boxes represent the discovered vector projection similarity condition between the source box and the target box. The representative element is the dark grey one.

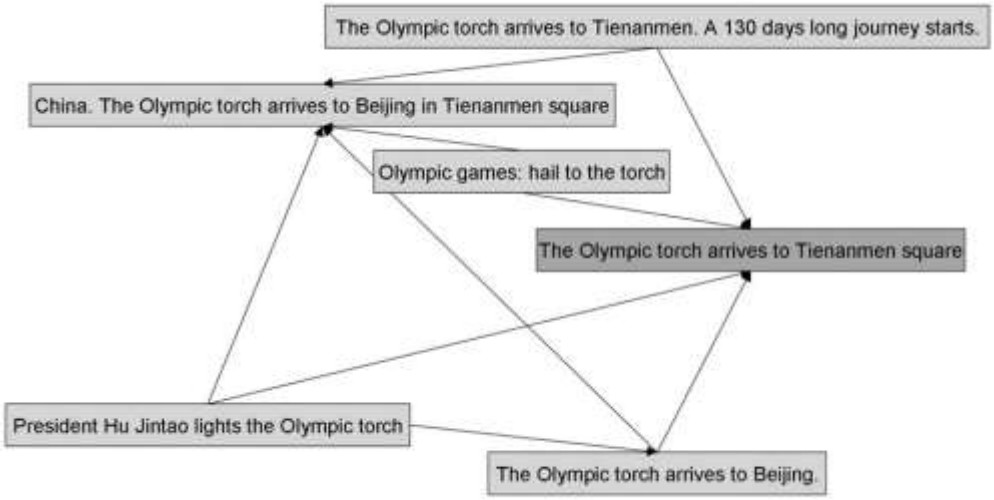

Figure 2. Example of vector projection similarity network.

C. Implementation of R

The implementation of the semantic relevance function has the following expression:

$$R(a, b) = \max\{I(q(L(s(a))), t(b)) - \zeta, 0\} \quad (4)$$

Where s(a) is the set of distinct sentences available, q() is a query constructor and t(b), Finally, I() is the relevance of t(b) w.r.t. q(L(s(a))) as calculated by the full text index, and ζ is an empirical parameter relaxing or restricting the minimum relevance value accepted. The output of the linguistic analysis L() is a vector of lemmas, each of which is tagged with its linguistic token (e.g. verb, noun, proper noun). The query constructor q() works as follows:

• From each sentence si e a (e.g. one title phrase and some description sentences), it selects lemmas tagged as nouns and builds an or-ed query fragment qi with them.

• It builds a complex query by joining all available query fragments qi, with a decreasing weight schema. The weighting schema associates higher weights to query fragments derived from sentences occurring earlier, so that to give more importance to the title and to the initial description sentences than to the middle or final ones (see example of Figure 3).

4. Experimental Results

To evaluate the overall effectiveness of our system we set up a pool of 25 expert users, taken from the RAI – Radio televisione Italiana employees, to whom we asked to make some evaluation sessions. Among the measured indicators, there were the following two:

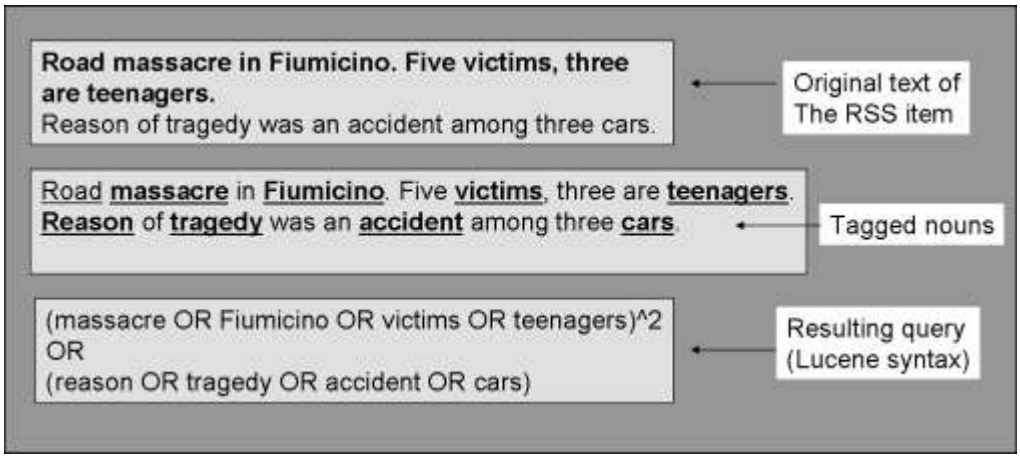

Figure 3. Example of linguistic analysis.

Average cohesion of the aggregations (α1); b) average relevance of the correctly selected titles (α6). Both indicators have been rated by users on a scale from 1 (poor) to 5 (excellent). Average values for α1 and α6 have been respectively 4.23 and 4.85. The demonstration will guide the guest through some of the basic and advanced features of the system, and in particular: a) a browsing interface on the database of detected and segmented newscast programs, (Figure 4); b) a navigation interface through the multimodal aggregations found by the system in form of a multimodal RSS feed (Figure 5); c) a search interface through which queries on a full text index of the found multimodal aggregations can be freely run at the guest's request.

Figure 4. Browsing the news items database

5. Conclusion

In our opinion, the significance of this contribution in the context of ICDM is manifold: a) it demonstrates a novel generic hybrid clustering method able to produce directed graphs out of the aggregated item sets, following a projection similarity measurement; b) it demonstrates the feasibility of a multimodal aggregator system component which applies the developed method in the case

of multimodal news item collection and delivery; c) it allows the guests to concretely touch the results of the system by providing full access to a full-text indexed database of thousands television news items and collected online articles.

References

1. F. Brugnara, M. Cettolo, M. Federico, and D. Giuliani. A system for the segmentation and transcription of italian radio news. In Proc. of RIAO, Content-Based Multimedia Information Access,2019.
2. Messina, R. Borgotallo, G. Dimino, D. Airola, and L. Boch. Ants: A complete system for automatic news programme annotation based on multimodal analysis. In Proc. Of WIAMIS 2018.
3. H. Schmid. Probabilistic part-of-speech tagging using decision trees. In Proc. of International Conference on NewMethods in Language Processing, 2018
4. Nandhini. R, Pavithra. P, Abinaya. P and S.Manikandan, "Information Technology Architectures for Grid Computing and Applications ", International Journal of Advanced Research Computer Engineering and Technology, ISSN:2278-1323, Vol.3, No.06, pp:2239-2242,June'2014.
5. Fran Berman, Geoffrey Fox and Tony Hey, "Grid Computing: Making the global infrastructure a reality", Published by Wiley,2018
6. Luis Ferreira and Non Keung, " Grid Computing Products and Services" by IBM Redbooks Publications, ISBN 0738491780,2013
7. Rich Wolski and Todd Bryan, "Grid Resource Allocation and Control Using Computational Economies", University of Tennessee, Wheaton College,2012
8. Ives et al Z.G (2015) "An Adaptive Query Execution System for Data Integration," Proc. ACM SIGMOD.
9. Baskins Judy Arrays D (2014) http://judy.sourceforge.net.
10. Bornea M.A , Vassalos V, Kotidis Y, and Deligiannakis A, (2019) "Double Index NEsted-loops Reactive join for Result Rate Optimization," Proc. IEEE Int'l Conf. Data Eng. (ICDE).
11. S Manikandan, K Raju, R Lavanya, R.G Gokila, "Web Enabled Data Warehouse Answer With Application", Applied Science Reports, Progressive Science Publications, E-ISSN: 2310-9440 / P-ISSN: 2311-0139, DOI: 10.15192/PSCP.ASR.2018.21.3.8487, Volume 21, Issue 3, pp. 84-87, 2018
12. Negri M. and Pelagatti G. (2017) "Join During Merge: An Improved Sort Based Algorithm," Information Processing Letters vol. 21, no. 1, pp. 11-16.
13. S. Manikandan, K. Raju, R. Lavanya, R.Hemavathi, "Energy Efficiency Controls on Minimizing Cost with Response Time and Guarantee Using EGC Algorithm",International Journal of Information Technology Insights & Transformations, Vol. 3, No. 1, 2017.