# A Generalised Cross-Modal Clustering Method Applied to Multimedia News Semantic Indexing and Retrieval

Alberto Messina*
RAI Radiotelevisione Italiana
Centre for Research and Technological Innovation
Corso Giambone 68, I-10135 Turin, Italy
a.messina@rai.it

Maurizio Montagnuolo
RAI Radiotelevisione Italiana
Centre for Research and Technological Innovation
Corso Giambone 68, I-10135 Turin, Italy
maurizio.montagnuolo@rai.it

## ABSTRACT

Current Web technology has enabled the distribution of informative content through dynamic media platforms. In addition, the availability of the same content in the form of digital multimedia data has dramatically increased. Content-based, cross-media retrieval applications are needed to efficiently access desired information from this variety of data sources. This paper presents a novel approach for cross-media information aggregation, and describes a prototype system implementing this approach. The prototype adopts online newspaper articles and TV newscasts as information sources, to deliver a service made up of items including both contributions. Extensive experiments prove the effectiveness of the proposed approach in a real-world business context.

## Categories and Subject Descriptors

H.0 [**Information Systems**]: General; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Linguistic processing*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Clustering*

## General Terms

Algorithms, Experimentation

## Keywords

Cross-modal clustering, multimedia mashup, news retrieval

## 1. INTRODUCTION

In recent years, the global diffusion of the Internet and the progress in developing Web multimedia applications are enabling the delivering of dynamic heterogeneous content such as news, blogs and audio/video podcasts. This content is commonly published through RSS feeds. Users can manage RSS feeds using a feed reader that periodically downloads the updated content from the subscribed feeds, displays the items in each feed and provides links to the related resources. However, the basic functionalities of these readers return the unorganised list of all the items included in the subscribed feeds, causing an information overload effect. Hence, effective solutions for intelligent information fusion and organisation are becoming indispensable. Here, the challenge lies in the ability of combining and presenting heterogeneous data coming from multiple information sources, i.e. *cross-modal*, and consisting of multiple types of content, i.e. *multi-media*.

This paper, a substantial extension of the original idea presented in [13], describes a framework for content-based, cross-media information aggregation, and its application to the real case of multimedia news retrieval. The paper is organised as follows. Section 2 reviews related work. Section 3 introduces a model of the domain. Based on this background, Section 4 describes our cross-modal clustering algorithm. Section 5 provides a description of our prototype architecture, and Section 6 details the core technologies used in its development. Section 7 presents the performance of the system. Finally, Section 8 provides conclusive remarks and future plans regarding the presented research.

## 2. RELATED WORK

### 2.1 Information Mashup

Information mashup is becoming a hot topic in the WWW community. A mashup is a Web application that aggregates content from different data sources to deliver a new, hybrid service that was not originally supported. Recently, many tools have been released for this purpose, such as Google Mashup[1], Yahoo! Pipes[2] and Microsoft Popfly[3].

Much of the current work involves grouping data from only a single domain and from only a single media, such as RSS items aggregated according to a taxonomy of concepts [12, 17]. As an RSS feed usually contains only short descriptions of the referenced items' content, the aggregation process may not be a trivial task. The works in [1, 6] employ either the user's interaction, or external knowledge sources, to improve the aggregation performance.

Nonetheless, the nature of the data to be aggregated cannot be in principle shrinked to be merely mono-media. Therefore, tools to integrate multi-media data from mono-modal information sources were investigated. A method for querying persons in Yahoo! News images using the enclosed news captions is presented in [7]. In [9], a speaker recognition

*and University of Turin, Department of Computer Science, Turin, messina@di.unito.it


*WWW 2009,* April 20–24, 2009, Madrid, Spain.
ACM 978-1-60558-487-4/09/04.

[1] http://editor.googlemashups.com/
[2] http://pipes.yahoo.com/
[3] http://www.popfly.com/

system based on facial, vocal and prosodic features is presented. In [18], a multimedia integration system to deliver personalised tourist Web services is proposed.

More challenging approaches are those employing both cross-modal information channels, like radio, TV and the Internet and multi-media data such as audio, video and text [19, 20]. More specifically, these approaches aims at enriching traditional TV broadcasts with semantic metadata derived from non-traditional information sources as the Internet. Similarly to our work, the authors in [3, 10, 15] address the problem of finding news articles on the Web relevant to the ongoing stream of TV broadcast news.

## 2.2 Topic Detection and Tracking

Additional works particularly relevant to our application domain are those pursued under the NIST Topic Detection and Tracking (TDT) project.[4] TDT aims at automatically locating, linking and accessing topically related information items within heterogeneous, real-time news streams. Intuitively, a *topic* is defined as an aggregation of information items that are *semantically relevant* to a real-world event. As an example, a earthquake could be the event that triggered the topic. Any information item, such as a newscast story or a newspaper article, that talks about the earthquake, or e.g. the rescue attempts, the number of casualties, and so on, is semantically relevant to the topic. The identified tasks of TDT are News Story Segmentation (NSS), First Story Detection (FSD), Topic Detection (TD), Link Detection (LD), and Topic Tracking (TT).

The news story segmentation task concerns the ability of automatically detecting semantically coherent parts of the news streams, such as a single news story. The TRECVID[5] initiative had news segmentation among its tasks in 2003 and 2004. The common base of the approaches for automatically segmenting TV newscasts into individual news stories consists in using a combination of visual, audio and speech features. Systems performance evaluation in the TRECVID news story segmentation task is presented in [2]. We perform the news story segmentation process by exploiting aural and visual cues, with the help of a three-layered heuristic framework inspired by the editorial rules used by newscasts producers, as it will be explained in Section 6.1.

The first story detection task is aimed at recognising information items linked to topics never seen before. FSD is typically approached by representing each information item as a set of features (e.g., newswire text or closed transcriptions of radio and TV speech). When an incoming item is received, its feature set is matched against those of all the past items according to a similarity measure. If, for each past item, the similarity measure is above a fixed threshold, then the incoming item is marked as new. Following the same approach, the topic and link detection tasks aim at aggregating and linking individual information items related to the same topic. The last task aims at keeping track of information items similar to a set of example items.

## 2.3 Contributions of This Paper

Despite the great number of research activities, the current state of the art techniques for the development of a complete multimodal (i.e., cross-modal and multi-media) framework are still far from satisfying expectation. Our work innovates in this direction, providing a methodology able to exploit the potentialities coming from the integration of heterogeneous information sources and media modalities. In particular, we adopt online newspaper articles and TV newscasts as information sources, to deliver multimodal search and retrieval services, integrating items coming from both contributions.

Closely related to our work are those presented in [3, 10, 15]. However, such approaches suffer from some limitations as they only provide one-way, one-to-many relationships, i.e. from single TV news stories to multiple Web pages. Instead, in our approach bidirectional, many-to-many relationships between TV and the Web are provided, thus augmenting the flexibility and versatility of the proposed framework.

Additional relevant works are those aimed at topic detection and tracking. The fundamental problem of almost the current TDT approaches lies in the definition of the similarity measure that is used to evaluate the distance between items. If the feature sets extracted from two information items are homogeneous from the data representation and data semantics perspectives, it is possible to define a similarity measure among them. Consequently, any clustering algorithm based on that similarity measure can be applied to discover aggregations of information items. Unfortunately, in many practical cases it happens that information items are not homogeneous, e.g. when text documents and multimedia objects have to be aggregated. While in this case it is possible to define similarity measures in *each* of the two spaces, it is difficult to establish a similarity measure in the hybrid space constituted by the *union* of the two spaces. A solution to this problem is provided by cross-modal (or hybrid) clustering [4]. Moreover, existing clustering methods typically output groups of items with no intrinsic structure. This inhibits the possibility of defining representative elements other than simple cluster centroids, as well as the ability of discovering deeper relations among grouped items like equivalence and entailment. In fact, the existing methods mostly link information items symmetrically, e.g. using the cosine similarity as distance metric [16]. This means that two linked items relate each other with the same strength. However, in most cases two related items should be assigned different strength to link each other. Our framework uses a cross-modal clustering algorithm whose kernel is an asymmetric relevance function between information items. The function asymmetry guarantees that different strength of relations can be discovered among information items.

Summing up, the innovation of our work can be presented in two aspect. First, we provide a general method for the aggregation of information streams based on the concept of semantic relevance and on a novel asymmetric aggregation function. Then, we present a fully unsupervised framework that implements all the functionalities provided by the general method. User experiments show the applicability and effectiveness of our solution in a real-world business context.

# 3. PROBLEM SETTING

This section presents a model of the information flow in the news publishing process. A typical news production and distribution cycle is shown in Figure 1. The cycle starts with the *news event* $e$, i.e. any relevant fact happened at some time and place, along with all the *information elements* generated during its occurrence. For example, the voting for a law or the presentation of a pe-

[4] http://www.nist.gov/speech/tests/tdt/

[5] www-nlpir.nist.gov/projects/trecvid/

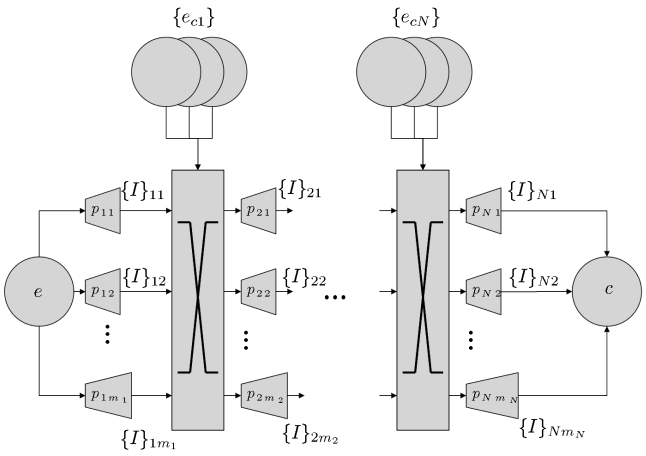


**Figure 1: Illustration of news publishing workflow. The occurrence of a real-world event $e$ generates many information items $I_{ij}$. Information providers $p_{ij}$, e.g. press agencies and broadcasters, merge and deliver these information items to the final consumers $c$ through multiple distribution channels, e.g the Internet, radio and TV, and presentation platforms as speech, video or text.**

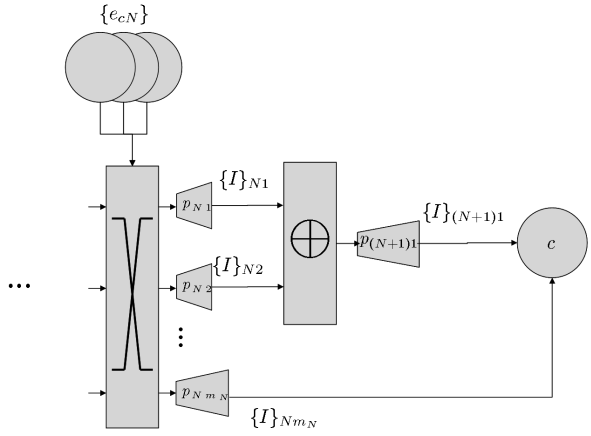


**Figure 2: Illustration of multimodal aggregation. Heterogeneous information streams $\{I\}_{N1}$, $\{I\}_{N2}$ sharing common semantics are aggregated and presented to the consumers in an integrated form.**

tition could be the information elements generated during a parliamentary session. The information elements are collected by a pool of *primary providers*, e.g. the reporters of a press agency, $\mathcal{P}' = \{p_{1j}\}, \ \forall j = 1, \ldots, m_1$, who in turn package them into *information items* $I_{1j}$ that represent coherent, finite and consumable information units, e.g. the press agency journalistic reports. These information items can be afterwards caught by the *secondary providers* $\mathcal{P}'' = \{p_{2j}\}, \ \forall j = 1, \ldots, m_2$, who can unpack them and use part of the contained elements to build new information items, optionally enriched by additional information derived from some contextual events $\{e_{ck}\}$. The process can be iterated until some editorial criteria is satisfied. For example, a news agency can pick up different journalistic reports and produce an article to be published on a newspaper. As a result, original information elements reach the *final consumer* $c$ through a variety of delivery paths, e.g. Web news, print and broadcast media, and packaged in distinct information items, e.g. RSS feeds, newspapers articles or newscasts.

The task of an aggregation information system is detecting and reconstructing a *surrogate* of the original event $e$ and of its most relevant contextual events, by combining the informative contributions coming from a subset of the information items sets in scope of the consumer $c$, i.e. a subset of $\{\{I\}_{N1}, \{I\}_{N2}, \ldots, \{I\}_{Nm_N}\}$. Since the nature of all the intermediate information providers and of the information elements cannot be in principle shrinked to be merely textual, we assume that information items may be multimedia, i.e. they can be presented in text, speech and visual formats. Figure 2 illustrates schematically the role of a component, denoted with $\oplus$, performing an aggregation process between two sources. This can be viewed as an intermediate process that merges contributions $\{I\}_{N1}$ and $\{I\}_{N2}$, delivered respectively by the providers $p_{N1}$ and $p_{N2}$, to generate a contribution $\{I\}_{(N+1)1}$ delivered to the consumers $c$ through the provider $p_{(N+1)1}$.

# 4. CROSS-MODAL CLUSTERING

Multimodal aggregation is performed by cross-modal clustering based on the concept of *semantic relevance*, which is inspired by the definition originally proposed in [11].

*Definition 1. (Semantic relevance).* Let $\pi$ and $\beta$ be two information items reached to the consumers through information streams $\{I\}_{N1}$ and $\{I\}_{N2}$, respectively. In this context, the *secondary* information item $\beta$ is semantically relevant to the *primary* information item $\pi$ if the fruition of $\beta$ by consumers satisfies the consumers expectations about $\pi$.

Semantic relevance is modelled by the linking function $R(\cdot)$ that measures how likely the secondary information items are relevant to the information needs expressed by the primary information items. Cross-modal clustering is able to discover these semantic relations in heterogeneous data, thus providing facilities to effectively retrieve desired information in cross-modal, multimedia information streams. The algorithm is detailed in the following subsections.

## 4.1 Affinity Analysis

Let $\Pi = \{\pi_i\}_{i=1}^{m}$ and $\mathcal{B} = \{\beta_j\}_{j=1}^{n}$ be two sets of information items, for which a distance metric in the space $\mathcal{H} = \Pi \cup \mathcal{B}$ is not defined. Let $R : \Pi \times \mathcal{B} \rightarrow [0, 1]$ be a linking function such that:

- $R(\pi_i, \beta_j) \rightarrow 1$ (tends to 1) if $\beta_j \in \mathcal{B}$ is semantically relevant to $\pi_i \in \Pi$;
- $R(\pi_i, \beta_j) \rightarrow 0$ (tends to 0) if $\beta_j \in \mathcal{B}$ is not semantically relevant to $\pi_i \in \Pi$.

Figure 3 shows an example in the context of Web and TV news. Many online news articles often deal with the same fact. In addition, these textual resources can be linked to a series of TV news reports, yet giving different updates or

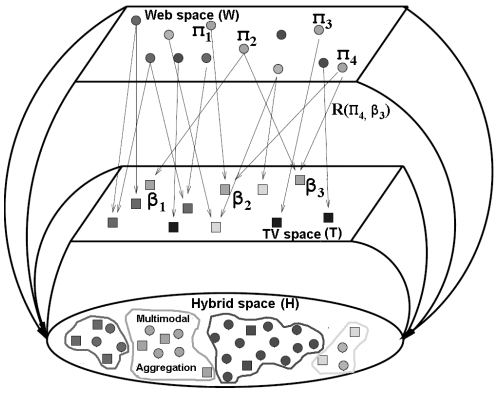


**Figure 3: Illustration of cross-modal clustering. Semantically relevant primary information items (e.g., Web assets $\pi_i$) and secondary information items (e.g., TV assets $\beta_i$) are merged in a new hybrid space $\mathcal{H}$, thus generating a multimodal aggregation.**

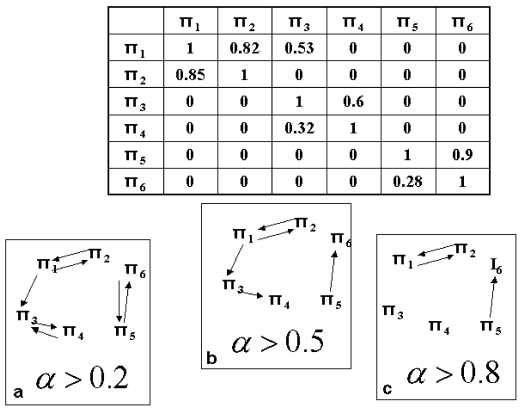


**Figure 4: Example of the equivalence matrix between primary information items $\{\pi_i\}_{i=1}^{6}$ and the corresponding connectivity graph for three values of $\alpha$.**

viewpoints on the same matter over time. Consumers should be made able to exploit these semantic links and to track sequels of linked stories in a new, hybrid space, including both the Web and TV contributions.

We define the *affinity matrix* as $\mathbf{A} = (\mathbf{r_1}, \ldots, \mathbf{r_m})^T \in [0,1]^{m,n}$, where

$$\mathbf{r_k} = (R(\pi_k, \beta_1), \ldots, R(\pi_k, \beta_n)), \quad k = 1, \ldots, m \qquad (1)$$

Intuitively, the construction of $\mathbf{A}$ can be seen as a *space transformation* process, which links the information items from the primary space $\Pi$ to the secondary space $\mathcal{B}$, according to the semantic relevance between objects in such spaces.

## 4.2 Hybrid Matching

Once the affinity matrix has been constructed, the similarity between primary information items $\pi_i, i = 1, \ldots, m$ is evaluated by exploiting their projection in the space $\mathcal{B}$.

Let $(\pi_a, \pi_b) \in \Pi$ be a couple of primary information items represented by the affinity vectors $(\mathbf{r}_a, \mathbf{r}_b)$, where $\mathbf{r}_a, \mathbf{r}_b$ are the corresponding rows in matrix $\mathbf{A}$. The similarity between $\pi_a$ and $\pi_b$ in the secondary space $\mathcal{B}$ is defined as follows:

$$S(\pi_a, \pi_b) = \frac{\langle \mathbf{r_a}, \mathbf{r_b} \rangle}{\|\mathbf{r_a}\|^2}\ , \qquad (2)$$

where $S(\cdot)$ is the *affinity vector similarity* function and $\|\cdot\|$ is the norm induced by the inner product in $[0,1]^n$.

Intuitively, the function $S(\cdot)$ defined in (2) measures *how much* the information item $\pi_a$ is explained by the information item $\pi_b$, in the space of their affinity vectors.

The function $S(\cdot)$ has the following properties:

$$S(\pi_a, \pi_b) = \cos(\mathbf{r_a}, \mathbf{r_b}) \frac{\|\mathbf{r_b}\|}{\|\mathbf{r_a}\|} \qquad (3)$$

$$\text{iff } S(\pi_a, \pi_b) > \alpha \ \wedge\ S(\pi_b, \pi_a) > \alpha \text{ then } Eq(\pi_a, \pi_b) \qquad (4)$$

$$\text{iff } S(\pi_a, \pi_b) > \alpha \ \wedge\ S(\pi_b, \pi_a) \leq \alpha \text{ then } Ent(\pi_a, \pi_b) \qquad (5)$$

where $\cos(\cdot)$ is the cosine similarity defined in $\mathcal{B}$, and $\alpha$ is a fixed threshold such that $\alpha \in [0,1]$. Equation (4) introduces the *semantic equivalence* relation between $\pi_a$ and $\pi_b$, $Eq(\pi_a, \pi_b)$, while Equation (5) introduces the *semantic entailment* relation from $\pi_a$ to $\pi_b$, $Ent(\pi_a, \pi_b)$. Notice that the latter relationship would not be discovered by using the plain cosine similarity measure. Intuitively, the asymmetry given by Equation 2, introduces the possibility to have hierarchies among the aggregated objects, providing also means for a natural procedure to discover representative elements and to have a multi-level granularity of presented information. The disadvantage w.r.t. symmetric measurements is the introduction of extra computation.

The affinity vector similarity function is computed for each couple of affinity vectors $(\mathbf{r_a}, \mathbf{r_b})$, $a, b = 1, \ldots, m$. The result is the equivalence matrix $\mathbf{E} = (e_{ab}) \in \Re^{m,m}$, where:

$$e_{ab} = \begin{cases} 1, \text{ if } a = b \\ S(\pi_a, \pi_b), \text{ if } a \neq b \text{ and } S(\pi_a, \pi_b) \geq \alpha \\ 0, \text{ otherwise.} \end{cases} \qquad (6)$$

## 4.3 Induced Partitions

Once $\mathbf{E}$ is calculated, the primary connectivity graph $G = (V, E)$ is built. Each node of the graph corresponds to a primary information item $\pi_i$. Two nodes $v_a, v_b$ are connected from $v_a$ to $v_b$ if the corresponding element $e_{ab} \in \mathbf{E}$ is greater than $\alpha$. The $\alpha$-cut value guarantees that every pair of linked information items has a semantic relevance of at least $\alpha$. Figure 4 shows an example.

Analysing the disconnected subgraphs included in $G$, a partition of the graph nodes $D(\alpha) = \{\gamma_1, \ldots, \gamma_{|D(\alpha)|}\}$ is built. $D(\alpha)$ is the partition of the primary space $\Pi$ induced by the space $\mathcal{B}$. For example, from Figure 4(a), it would be

$$D(\alpha) = \{\gamma_1, \gamma_2\} = \{(\pi_1, \pi_2, \pi_3, \pi_4), (\pi_5, \pi_6)\},$$

with $\alpha > 0.2$. Each part $\gamma_i \in D(\alpha)$ constitutes a set of semantically related primary information items linked according to their semantic relationships. The parameter $\alpha$ defined in (6) governs the structure of the resulting partition, by relaxing ($\alpha \downarrow$) or restricting ($\alpha \uparrow$) the conditions under which the elements of $\Pi$ can be aggregated.

## 4.4 Representative Elements

For each part $\gamma_i \in D(\alpha)$, its representative element $\bar{\pi}_i$ is chosen so that:

$$\bar{\pi}_i = \arg \max_{\pi_{ij} \in \gamma_i} \sum_{k \neq j} S(\pi_{ik}, \pi_{ij})\ . \qquad (7)$$

Equation (7) means that the representative element is that one whose affinity vector total similarity measurement is maximised. This criterion is based on the observation that the higher is the affinity vector total similarity, the higher is the number of elements in the aggregation that are semantically entailed by it, so that the item content is expected to be the most complete w.r.t. the semantics of the partition, and therefore the more representativeness is conveyed by the item itself. To stick with the example shown in Figure 4(a), the representative element for the part $\gamma_1 = (\pi_1, \pi_2, \pi_3, \pi_4)$ would be $\bar{\pi} = \pi_1$.

## 4.5 Multimodal Aggregations

Given an induced partition $D(\alpha)$ we can finally build the set of multimodal aggregations $D(\alpha)^* = \{\gamma_1^*, \ldots, \gamma_{|D(\alpha)|}^*\} \subseteq 2^{\mathcal{H}}$ by retrieving the elements of $\mathcal{B}$ relevant to each $\gamma_i \in D(\alpha)$, as follows (with $K = |D(\alpha)|$):

$$\forall i : \gamma_i^* = \gamma_i \cup B_i, \quad i = 1, \ldots, K \tag{8}$$

$$B_i = \cup_{j=1}^{N_j} \beta_{ij} \tag{9}$$

$$\beta_{ij} = \{b \in \mathcal{B} : R(\pi_{ij}, b) > \eta\} \ , \tag{10}$$

where $\eta$ is a parametric threshold. The function of $D(\alpha)^*$ is that of integrating the partition $D(\alpha)$ with the semantically relevant elements of $\mathcal{B}$. Notice that $D(\alpha)^*$ is not in general a partition of $\mathcal{H} = \Pi \cup \mathcal{B}$, because elements of $\mathcal{B}$ may be semantically relevant to elements of $\Pi$ belonging to different elements of $D(\alpha)$, and because some elements of $\mathcal{B}$ may be not semantically relevant to any element of $\Pi$.

# 5. PROTOTYPE ARCHITECTURE

We applied our cross-modal aggregation framework to a concrete case: clustering Web and TV news streams. The prototype architecture is shown in Figure 5. The system is a processing machine having two inputs, i.e. digitised broadcast news streams (DTV) and online newspapers feeds (RSSF), and one output, i.e. the multimodal aggregation service (MMAS), that is automatically determined from the semantic aggregation of the input streams. In connection with the model presented in Section 3, the prototype is thus a particular implementation of the general architecture shown in Figure 2, where RSSF and DTV are the two information streams $\{I\}_{N1}$ and $\{I\}_{N2}$.

## 5.1 DTV Stream Input Chain

The complexity introduced by considering digital television as an information source is primarily constituted by the necessity of automatically detecting and identifying information items in the real-time acquired stream. For this purpose, in our system the DTV stream is at first analysed and partitioned into programmes using a visual pattern matching algorithm. Video elements (shots) indicating starting and ending of programmes are used as reference prototypes to be searched through the acquired video stream [14]. On such detected programmes, automatic segmentation into elementary news stories is performed, as it will be presented in Section 6.1. Once segmented, the audio track of each story is analysed by an automatic speech recogniser tool [5], providing text transcriptions of the spoken parts in the DTV stream. Both English and Italian languages are supported. Finally, the detected news stories are indexed in the TVi documents catalogue. Summing up, the output of the DTV stream processing chain is an index structure,

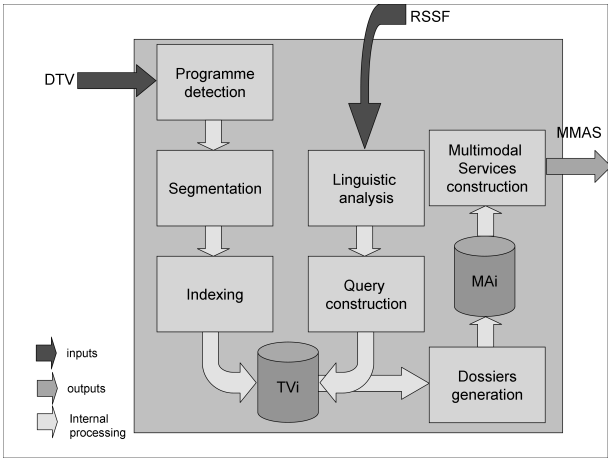


**Figure 5: Functional system architecture.**

whose contents are the news stories automatically detected by processing broadcast TV programmes.

## 5.2 RSSF Stream Input Chain

The RSSF stream consists of RSS feeds from several major online newspapers and press agencies. Additionally, also users weblogs can be treated. Quite similarly to the official information sources represented by online newspapers, users weblogs can be used to build the aggregations, provided that they are published and delivered as RSS feeds.

On each RSS item, a linguistic analysis is performed to identify meaningful linguistic structures, e.g. verbs, nouns, adjectives, within the RSS items' content. This information is used to build a set of lexical terms that capture the semantic concepts expressed in the articles linked by the RSS feeds. The technical details underlying the functional interface of the RSSF processing chain are described in Section 6.2.

The outputs of the linguistic analysis are then employed by the query constructor to generate a set of representative query expressions, which are submitted to the index structure of the TVi documents catalogue. For each item, the result of this search operation is a weighted set of newscasts stories of decreasing affinity to the target query.

## 5.3 MMAS Stream Output Chain

The results of the queries on the TVi index structure are used by the cross-modal clustering process to aggregate information items based on their semantic similarities, as previously described in Section 4. These aggregations are indexed and stored in the multimodal aggregation index (MAi). For each aggregation, the MAi stores the list of the aggregated RSS items and news stories, as well as a text document including the RSS itemsŠ titles and description phrases and the news stories transcriptions constituting the aggregation.

Operationally, the MAi is a persistent data repository from which the multimodal services are delivered to the users. The services currently supported by the system are outlined in the following subsections.

### *5.3.1 Multimodal Navigation*

Multimodal navigation is the ability of providing links between heterogeneous information items. Here, the users are able to browse the lists of broadcast news stories (i.e., the

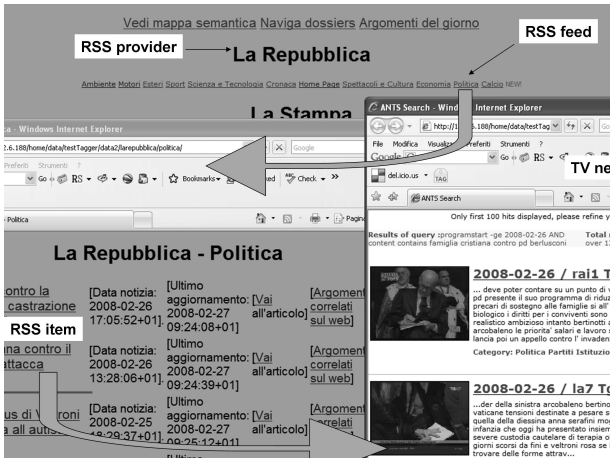


**Figure 6: Example of multimodal navigation.**

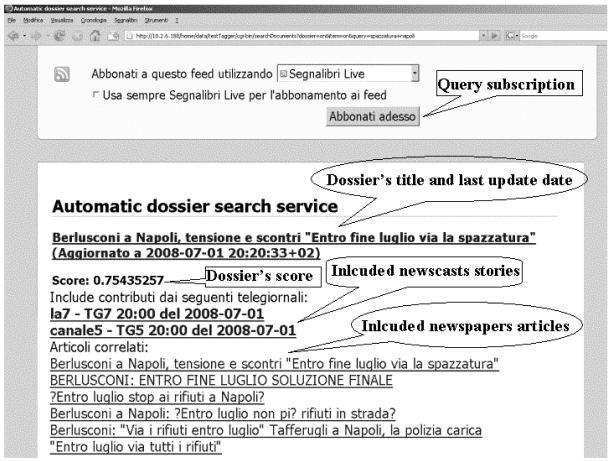


**Figure 7: Example of multimodal search & retrieval.**

target elements) related to the RSS items (i.e., the source elements). In this manner, the target elements are contextualised by the source elements. Thus, a context-guided browsing of cross-modal and multi-media (i.e., multimodal) content is offered to the users, as shown in Figure 6.

### *5.3.2 Multimodal Search and Retrieval*

The system supports both simple queries (e.g., one or more search keywords) as well as more advanced queries (e.g., weighted queries, boolean operators) for searching and retrieving the aggregations. As a simple example, Figure 7 shows the first result for the query *"garbage AND Naples"*.

To facilitate the results visualisation, the system provides a browsable Web page showing the ranked results. For each retrieved aggregation (also called "dossier"), the system lists the basic information, i.e. title, score and update time, and provides the links for the included news stories and newspaper articles.

In addition, as the search results are provided in the form of RSS feeds, users can subscribe to the submitted query, and automatically receive a notification when the results page is modified, i.e. when either an already included aggregation is updated or a new one is discovered.

### *5.3.3 Multimodal User Recommendation*

The system provides a recommendation service that helps users find the desired information according to their behavior and interests. The delivering of the service employes the queries submitted by a user to build a list of related queries. These system-generated queries can be then issued by the user to tune or redirect the search process. The query generation process is based on the assumption that the relevant aggregations for a query $q$ share some terms apart from the original terms used in $q$. Details of how this query derivation process works are given in Section 6.4.

# 6. CORE TECHNOLOGIES

This section describes the core technologies used to implement our prototype system.

## 6.1 Newscasts Detection and Segmentation

Segmentation of broadcast TV streams into programmes is performed by adopting an optimised video clip matching technique. A set of feature signatures are extracted from each frame of the acquired video clips, including colour, texture and motion histograms. Among all those extracted, features are selected that maximise the statistical divergence w.r.t. a sample population of the event to be searched, e.g. a programme's jingle. The selected feature vectors are then matched against those indicating starting and ending of programmes, using the histogram intersection distance.

Once detected, TV newscasts are automatically segmented into their elementary news stories. The segmentation process is done exploiting aural and visual cues with the help of a three layered heuristic framework. The used heuristics are based on the editorial rules employed by the TV editors in the news production process.

The basic heuristic $H_1$ considers boundaries of shots containing the anchorman as equivalent to news stories boundaries. The anchorman shots are detected using a second heuristic $H_2$ that considers the most frequent speaker as the anchorman. This heuristic allows to select the speaker who most likely is the anchorman, provided that a speaker clustering process labels all the speakers present in the programme and associates them to temporal segments of the content.

As the application of $H_1$ and $H_2$ is not yet enough to discern situations where e.g. the anchorman presents several consecutive brief stories without interruptions filled with external contributions, or where the beginning of a story does not correspond to an anchorman shot, we employ a third heuristic $H_3$ based on the observation that the introduction of a new brief story is often accompanied by a camera shot change, e.g. from a close up shot to a wider one. Thus, to optimise the accuracy of segmentation, we use an adaptive threshold shot detection algorithm based on the displaced frame difference (DFD) computed on luminance samples of contiguous video frames. Adaptivity is based on the local statistics of the DFD, so that content having higher DFD variance is processed against higher thresholds.

Once the shot detection is completed, similar shots, i.e. those sharing similar visual content, are grouped together through a shot clustering process [14]. This allows us to detect and classify shot clusters as pertaining to studio shots containing the anchorman following the same frequency heuristic used for detecting the candidate speaker ($H_2$). This double clustering process (both on audio and on video) enables a simple and effective algorithm for speaker tracking

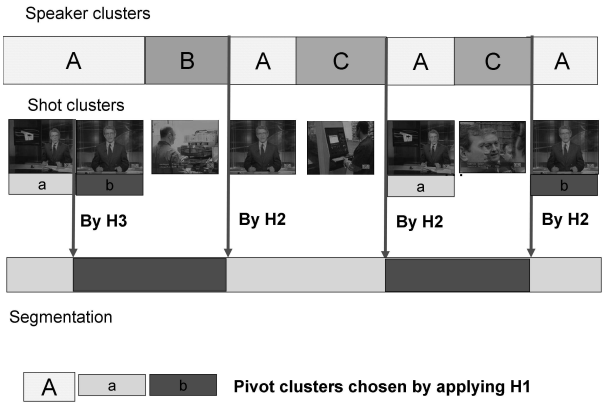


**Figure 8: Illustration of news story segmentation.**

and news story segmentation. Figure 8 illustrates an example. The anchorman shots $a$ and $b$ are detected according to the heuristic $H_2$ because both contain the same speaker $A$. As a shot boundary is detected between the shots $a$ and $b$, the first two stories are segmented according to $H_3$. The succeeding stories are then detected according to $H_1$.

Once detected, the spoken text of each news story is categorised according to its main topic using the `AI::Categorizer` framework,[6] and indexed by Lucene.[7] More detail on the whole programme segmentation process is provided in [14].

## 6.2 RSS Stream Processing

RSS streams are analysed to get the list of included items. Each item is represented by the tuple

$$\ell = (uuid, pubDate, link, title, description),$$

where *uuid* is a identifier that univocally identifies the item, *pubDate* is the publication date, *link* is the URL of the related online newspaper article, *title* is the headline of the corresponding article and *description* is a short summary of the corresponding article's content.

We split the title and the description fields into elementary phrases and tokenise them into words by applying the Tree Tagger tool[8] that labels each word according to a taxonomy of grammar terms, e.g. conjugating verbs, proper nouns, adjectives. Thus, an RSS item $\pi$ is represented by a vector of key/value pairs

$$\pi = ((k_{1t}, v_{1t})_{t=1}^{T}, (k_{2f}, v_{2f})_{f=1}^{D_1}, \ldots, (k_{ml}, v_{ml})_{l=1}^{D_{m-1}}), \quad (11)$$

where the keys are the words in the phrases, the values are the corresponding grammar terms, $T$, $D_1$ and $D_{m-1}$ are, respectively, the total number of tagged words in the title, in the first and in the last description phrase. The use of $\pi$ allows to extract elements of the title and description sentences that are important from the *linguistic* point of view, in opposition to statistical approaches (e.g., TF/IDF techniques) that simply rely on term frequency metrics, thus

[6] `http://search.cpan.org/dist/AI-Categorizer/`
[7] `http://lucene.apache.org/`
[8] `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`

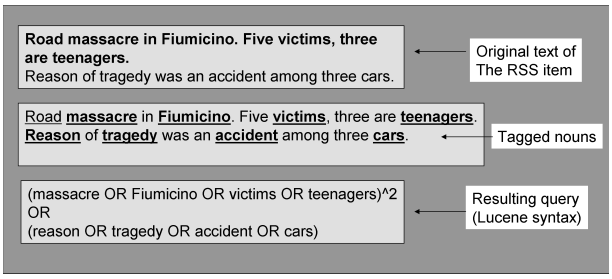


**Figure 9: Example of query construction.**

better simulating the human understanding of the semantic implied in the interpretation of short texts.

The vector $\pi$ is then used to generate a full text query string $Q$. The query construction process works in two steps. First, for each subvector $\mathbf{s}_i$ of $\pi$, an elementary query $q_i$ is built, selecting the words in $\mathbf{s_i}$ tagged as either *common noun* or *named entity* or *adjective*. Then, a combined query $Q$ is generated, joining all the elementary queries as follows:

$$Q := \bigcup_{i=1}^{m} q_i^{w_i} \quad (12)$$

$$w_i = 2^{(m-i)}, \quad \forall i = 1, \ldots, m\ . \quad (13)$$

This weighting schema associates higher weights to queries derived from phrases occurring earlier, in order to emphasise the title and the initial description phrases. An example of the query construction process is illustrated in Figure 9.

## 6.3 RSS Items and News Stories Aggregation

For each RSS item $\pi_i$ the associated query $Q_i$ is launched on the set of the broadcast news speech transcriptions indexed in the TVi (see Figure 5). The output of $Q_i$ is stored in the affinity results vector $\mathbf{r_i} = (r_{ij})_{j=1}^{n}$, where $r_{ij}$ is the query score of the news story $\beta_j$ to $\pi_i$. The set of all the affinity results vectors is then arranged in the affinity matrix $\mathbf{A}$ defined in Equation (1). From $\mathbf{A}$, we compute the equivalence matrix $\mathbf{E}$ of Equation (6), and build the correspondent connectivity graph $G$. We then proceed to discover the induced partition following the process presented in Section 4.3, and select the representative element of each part of the partition as described in Section 4.4. We construct the multimodal aggregations as defined in Section 4.5 cutting off elements for which $r_{ij} < \eta = 0.5$. A text document is generated including the titles, description phrases and transcriptions of the RSS items and news stories constituting the aggregation. Finally, all such documents are indexed by Lucene and made accessible through the MAi repository.

## 6.4 Derived Queries Generation

As introduced in Section 5.3.3, our system implements a user recommendation functionality through a query expansion mechanism. The expansion of user queries algorithm works as follows. Let $Q$ be a query submitted by the user $u$, and $\mathcal{A} = \{\gamma_i^*\}_{i=1}^{|A|}$ be the set of multimodal aggregations retrieved from the MAi (see Figure 5) for $Q$. For each aggregation $\gamma_i^* \in \mathcal{A}$, a feature vector $\mathbf{v} = (\mathbf{s}, \mathbf{c}, \mathbf{p})$ is extracted from the analysis of the RSS items' sentences (titles and description phrases), the referenced news articles text, and the TV news items' transcribed speech content. The sub-vector $\mathbf{s}$ stores the fraction of word occurrences in the aggregation,

according to a reference dictionary. The sub-vector $\mathbf{c}$ stores the normalised (w.r.t. the total number of objects in the aggregation) scores of the categories to which the aggregation belongs, according to the same set defined for the news story categorisation (Section 6.1). The sub-vector $\mathbf{p}$ is the set of couples of the proper nouns found by Tree Tagger in the RSS items included in the aggregation $\gamma_i^*$, and their corresponding frequencies.

The k-means clustering algorithm is run on the set $\mathcal{A}$ using $\mathbf{v}$ as feature vector, until either the desired precision $\epsilon$ is achieved, or the maximum number of epochs $N_{iter}$ is reached. Because of the heterogeneity of the sub-vectors of $\mathbf{v}$, we used the Euclidean distance to compare the sub-vectors in the $\mathbf{s}$ and $\mathbf{c}$ space, and the Jaccard distance to compare the sets in the $\mathbf{p}$ space. Given $\mathbf{v}_a = (\mathbf{s}_a, \mathbf{c}_a, \mathbf{p}_a)$ and $\mathbf{v}_b = (\mathbf{s}_b, \mathbf{c}_b, \mathbf{p}_b)$ two feature vectors, we define a combined distance used by the k-means clustering process:

$$d(\mathbf{v}_a, \mathbf{v}_b) = \frac{1}{3}\left(L2(\mathbf{s}_a, \mathbf{s}_b) + L2(\mathbf{c}_a, \mathbf{c}_b) + \frac{|\mathbf{p}_a \cap \mathbf{p}_b|}{|\mathbf{p}_a \cup \mathbf{p}_b|}\right) \quad (14)$$

Once the clustering process is completed, we select the centroid of the most populated cluster, $C_M = (\mathbf{s}_M, \mathbf{c}_M, \mathbf{p}_M)$ and select the proper nouns $p_1, \ldots, p_K$, $p_i \in \mathbf{p}_M$, such that the corresponding frequencies are greater than a dynamic threshold calculated as the mean of all frequencies in $\mathbf{p}_M$. We then derive the two queries $Q \wedge \{p_1 \ldots p_K\}$ and $Q \vee \{p_1 \ldots p_K\}$. Let us to consider the following example. Suppose a user submits the query *"Donadoni contratto"* (i.e., Donadoni's contract), presumably to find information about the contract of Roberto Donadoni. Let us suppose that the described clustering and selection process discovers the following proper nouns: *Abete*, *Federcalcio* (i.e., football federation), *Lippi* and *Marcello*. Then, the following derived queries would be proposed to the user:

$q_1$ := `(lippi abete marcello federcalcio)` $\wedge$ `(donadoni contratto)`, i.e. a refinement of $Q$;
$q_2$ := `(lippi abete marcello federcalcio)` $\vee$ `(donadoni contratto)`, i.e. an expansion of $Q$.

# 7. EXPERIMENTAL EVALUATIONS

This section presents the experimental evaluation of each part of our system. The system was run from the end of November 2007 to the beginning of June 2008. In this period of time, we collected about 88,280 online articles and 23,940 news stories, resulting from the segmentation of 3,670 newscast programmes. The online articles were downloaded from 95 RSS feeds supplied by 16 online newspapers and press agencies Web sites. The newscasts were acquired from the daily programming of seven national TV channels. We set $\alpha = 0.8$ in Equation (6), thus obtaining a total of 4,187 automatically generated aggregations.

## 7.1 Performance of DTV Stream Processing

### 7.1.1 Processing Time of Programme Detection and Segmentation

The processing conditions are characterised by intense bursts of activities concentrated around the main editions of the newscasts (around 2 pm and around 8 pm). The system acquires 7 major national channels 24 hours/day and 365 days/year, and elaborates 16 programmes/day approximately 8 for each burst period (total approximately 10 hours/day elaborated material). To accommodate these requirements, the system has been implemented on a distributed multi-CPU architecture. The programme segmentation task takes on average $\approx 3.74$ times the programme duration, that is normally a newscast of 30 minutes.

### 7.1.2 TV Programme Detection

Two distinct experiments were performed to test the programme boundary detection accuracy. The first was aimed at identifying 11 different reference clips from a data set of 782 clips randomly acquired from daily television schedules. The second consisted in detecting the starting and ending jingles of seven distinct news programmes (total 14 clips) in real-time, broadcast streams. In the first experiment, the achieved precision and recall were $\approx 0.80$ and $\approx 0.87$, respectively. In the second experiment the reached precision and recall were, respectively, 1.00 and $\approx 0.90$.

### 7.1.3 News Story Segmentation and Categorisation

In order to measure the quality of news segmentation we used an alignment measurement that takes into account starting and ending boundaries with different weights. In addition, it considers under-segmentation effects (i.e., when the detected story starts/ends after/before the actual story) as being more penalising than over-segmentation effects (i.e., when the detected story starts/ends before/after the actual story) on the measurement [8]. The system was tested against a test set of 84 programmes (i.e., $\approx 40$ hours of material) achieving a precision of 0.76 and a recall of 0.73. The news story subject categorisation task was performed using a naive Bayesian classifier. A data set of 25,000 automatic speech transcriptions was collected. The classifier was trained on four fifths of the available data using a standard subject taxonomy of 21 categories. The remaining data were used for testing, reaching an accuracy value of $\approx 0.82$.

## 7.2 Performance of MMAS Services

### 7.2.1 Multimodal Aggregation Efficiency

Let $\mathcal{P} = \{\gamma_i^* = \gamma_i \cup B_i\}_{i=1}^{|\mathcal{P}|}$ be a set of multimodal aggregations, and let $t_i$ be the title automatically assigned to $\gamma_i^*$. To test the overall efficiency of the multimodal aggregation service, we set up a pool of 25 users, taken from the employers of our organisation, who were unaware of the rationales of the system. Each user was asked to perform evaluations in front of an optimised evaluation interface, designed and implemented on purpose. The interface shows a random list of aggregations. Each aggregation was evaluated through the following markers, using a judgement scale from 1 (i.e., disappointment) to 5 (i.e., full satisfaction):

1. For each aggregation $\gamma_i^*$ assign a cohesion index $\Gamma_i$ that reflects the overall consistency of the multimodal aggregation.
2. For each of the aggregated RSS items $\pi_{ij} \in \gamma_i$, $j = 1, \ldots, |\gamma_i|$, assign a consistency index $\rho_{ij}$ to the concept expressed by the multimodal aggregation $\gamma_i^*$;
3. For each of the aggregated news stories $\beta_{ik} \in B_i$, $k = 1, \ldots, |B_i|$, assign a consistency index $r_{ik}$ to the concept expressed by the multimodal aggregation $\gamma_i^*$;
4. For each aggregation $\gamma_i^*$ choose a title $T_i$ among those belonging to the RSS items $\pi_{ij} \in \gamma_i$, and assign a representativity index $\tau_i$ to it;

**Table 1: Multimodal Aggregation Efficiency Indices**

| **Index** | $\alpha_1$ | $\alpha_2$ | $\alpha_3$ | $\alpha_4$ | $\alpha_5$ | $\alpha_6$ | $\alpha_7$ |
|---|---|---|---|---|---|---|---|
| **Score** $\mu$ | 4.23 | 4.65 | 4.24 | 4.66 | 1.55 | 4.85 | 4.63 |
| **SD** $\sigma$ | 0.85 | 0.89 | 1.17 | 0.70 | - | 0.62 | 0.77 |
| **CI** $\Delta_{start}$ | 4.19 | 4.62 | 4.20 | 4.59 | - | 4.84 | 4.47 |
| **CI** $\Delta_{end}$ | 4.35 | 4.68 | 4.28 | 4.70 | - | 4.89 | 4.69 |

The following performance indices can be then defined:

- Cohesion of the multimodal aggregations:

$$\alpha_1 = \frac{1}{|\mathcal{P}|}\sum_{n=1}^{|\mathcal{P}|}\Gamma_n \tag{15}$$

- Consistency of the Web and TV aggregations:

$$\alpha_2 = \frac{1}{|\mathcal{P}|}\sum_{n=1}^{|\mathcal{P}|}\frac{1}{|\gamma_n|}\sum_{m=1}^{|\gamma_n|}\rho_{nm} \tag{16}$$

$$\alpha_3 = \frac{1}{|\mathcal{P}|}\sum_{n=1}^{|\mathcal{P}|}\frac{1}{|B_n|}\sum_{m=1}^{|B_n|}r_{nm} \tag{17}$$

- Title representativity of the multimodal aggregations:

$$\alpha_4 = \frac{1}{|\mathcal{P}|}\sum_{n=1}^{|\mathcal{P}|}\tau_n \tag{18}$$

- System-settled and user-defined title agreement:

$$\alpha_5 = 5\frac{|\{\gamma_i^* \in \mathcal{P} : t_i = T_i\}|}{|\mathcal{P}|} \tag{19}$$

- Relevance of the correctly and wrongly selected titles.

$$\alpha_6 = \frac{\sum_{\gamma_i^*\in\mathcal{P}:t_i=T_i}\rho_{iR(t_i)}}{|\mathcal{P}|} \tag{20}$$

$$\alpha_7 = \frac{\sum_{\gamma_i^*\in\mathcal{P}:t_i\neq T_i}\rho_{iR(t_i)}}{|\mathcal{P}|}, \tag{21}$$

  where $R(t_i)$ is a function returning the index of the RSS item $\pi_i$ that was taken as the representative for the aggregation $\gamma_i^*$.

Indices $\alpha_1$ to $\alpha_4$ represent the mean values of their respective elementary markers. $\alpha_5$ counts the fraction of aggregations for which the title settled by the system agreed with the title chosen by the user. $\alpha_6$ (analogously $\alpha_7$) indicates on average how well the titles correctly (wrongly) settled by the system explain the topics of the assessed aggregations.

The aggregations were evaluated from March to June 2008, getting a set of 651 assessments. Table 1 reports the score, the standard deviation (SD) and the 95% confidence interval (CI) for each of the seven efficiency indicators. The full scale value is 5 for all indicators. The reported values are statistical indices that take into account subjective assessments as over- and under-voting. Thus, they can be used as measures of the overall efficiency of the MMAS service.

The performance of the aggregation algorithm is influenced by the effectiveness of the news segmentation process. In some cases, due to undersegmentation of the news stories, the cohesion of the aggregations decreases. However, the system shows an outstanding performance, getting a global efficiency index (i.e., the mean of indicators $\alpha_{1,\ldots,7}$) of 4.12 (over 5). This value is mainly negatively affected by the indicator $\alpha_5$, seeming to indicate that the algorithm used to choose the aggregations' titles should be further improved. However, the index $\alpha_6$ indicates that the titles automatically settled as representative derive from RSS items that were scored as very relevant to the aggregation concept. In fact, since $\mu_6 > \mu_4$, $\sigma_6 < \sigma_4$ and $\Delta_6 \cap \Delta_4 = \emptyset$ we can state that the titles settled by the system are more representative than the average, and that there is a higher level of agreement among reviewers about their relevance score. Furthermore, as $\mu_7 \simeq \mu_4$, it can be concluded that even if the title automatically settled is wrong, it still significantly explains the topic of the corresponding aggregation.

### 7.2.2 Multimodal Search & Retrieval Efficiency

The efficiency of the multimodal search and retrieval service was evaluated using the mean average precision (MAP). Let $\mathcal{Q} = \{q_k\}_{k=1}^{N}$ be the set of user-generated queries and $\mathcal{H}_k = \{h_{ik}\}_{i=1}^{R_k}$ be the set of retrieved documents for $q_k$, ranked according to the Lucene score. Average precision (AP) is the average of the precision scores at the ranks where relevant hits (w.r.t. the original query $q_k$) occur. AP depends on how the relevant hits are ordered in $\mathcal{H}_k$. In the best case all the relevant hits appear before any non-relevant ones, thus resulting in $AP_k = 1$. The mean average precision is the mean of AP over the full set $\mathcal{Q}$:

$$MAP = \frac{1}{N}\sum_{k=1}^{N}AP_k = \frac{1}{N}\sum_{k=1}^{N}\frac{1}{R_k}\sum_{i=1}^{R_k}g_k(i)p_k(i)\ , \tag{22}$$

where $g_k(i)$ is a binary function that returns 1 if $h_{ik}$ is relevant to $q_k$, and $p_k(i)$ is the precision after $i$ hits of $\mathcal{H}_k$.

In the experiments, users were asked to submit some queries to the system, and then mark each retrieved aggregation as relevant or not to the submitted query. According to TREC specifications, we evaluated 50 queries, achieving a MAP of 0.79. This proves that the proposed approach, namely fusing contributions coming from television and the Web into a single document to be indexed, enables the delivering of an effective search and retrieval service.

### 7.2.3 Derived Query Generation Efficiency

Analogously to multimodal aggregation efficiency, we used users' ratings to evaluate our query derivation method. Let $\mathcal{Q}^* = \{q_j^*\}_{j=1}^{|\mathcal{Q}^*|}$ be the set of derived queries from the set of original queries $\mathcal{Q}$. For each $q_j^* \in \mathcal{Q}^*$, we calculate:

1. Average precision $AP_j^*$ w.r.t. the set of retrieved objects for $q_j^*$;

2. Relevance degree $\rho_j$ w.r.t. the original query $q \in \mathcal{Q}$ from which $q_j^*$ derives, expressed by a score from 1 (i.e., "totally unrelated") to 5 (i.e., "completely related").

The following performance markers can be then defined:

- Mean average precision of the derived queries:

$$\alpha_8 = MAP^* = \frac{1}{|\mathcal{Q}^*|}\sum_{j=1}^{|\mathcal{Q}^*|}AP_j^* \tag{23}$$

- Relevance of the derived queries to the original queries:

$$\alpha_9 = \frac{1}{|q_j^* \in \mathcal{Q}^*|} \sum_{\rho_j \neq 0} \rho_j \quad (24)$$

- Effectiveness of the query derivation system:

$$\alpha_{10} = \alpha_8 \frac{\alpha_9}{5} \ . \quad (25)$$

Indices $\alpha_8$ and $\alpha_9$ are the mean values of their respective elementary markers. $\alpha_{10}$ measures the effectiveness of the query derivation process w.r.t. the initial topics of interest to the users. We set $k = 64$, $\epsilon = 0.05$ and $N_{iter} = 20$ for k-means. A set of 140 derived queries produced by the user panel was evaluated, getting the following results: $\alpha_8 = 0.85$, $\alpha_9 = 4.3$ and $\alpha_{10} = 0.73$. The results show that the use of the derived queries improves the number of relevant documents retrieved at the top of the results list. Additionally, high relevance to the original query is provided. Therefore, the method is helpful in finding new relevant documents for the users who formulated the original query.

## 8. CONCLUSIONS

In this paper we presented a novel methodology to support the delivery of multimodal aggregation services. Multimodality is the capability of fusing and presenting heterogeneous data, such as audio, video and text, from multiple information sources, such as the Internet and TV. The method is based on: (i) a semantic relevance function acting as a kernel to discover the semantic affinities of heterogeneous information items, and (ii) an asymmetric vector projection model on which semantic dependency graphs among information items are built and representative elements of these graphs can be selected. To prove the applicability of our technique, we developed a system for aggregating and retrieving online newspaper articles and broadcast news stories. Obtained results are very encouraging and demonstrate the robustness and effectiveness of the proposed method. Future work will focus on a comparative analysis on clustering performance using symmetric similarity functions, and on the optimisation of the programme segmentation algorithms. Additional work will explore the integration of further information sources such as images and radio data, and the use of more sophisticated query derivation models.

## 9. REFERENCES


[1] J. W. Ahn, P. Brusilovsky, J. Grady, D. He, and S. Y. Syn. Open user profiles for adaptive news systems: help or harm? In *Proc. of WWW07*, pages 11–20, 2007.

[2] J. Arlandis, P. Over, and W. Kraaij. Boundary Error Analysis and Categorization in the TRECVID News Story Segmentation Task. In *Proc. of CIVR05*, pages 103–112, 2005.

[3] R. Basili, M. Cammisa, and E. Donati. RitroveRAI: A Web Application for Semantic Indexing and Hyperlinking of Multimedia News. In *Proc. of the Intl. Semantic Web Conf.*, pages 97–111, 2005.

[4] L. Bolelli, S. Ertekin, D. Zhou, and C. L. Giles. K-svmeans: A hybrid clustering algorithm for multi-type interrelated datasets. In *Proc. of Web Intelligence 2007*, pages 198–204, 2007.

[5] 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*, 2000.

[6] A. S. Das, M. Datar, A. Garg, and S. Rajaram. Google news personalization: scalable online collaborative filtering. In *Proc. of WWW07*, pages 271–280, 2007.

[7] K. Deschacht and M.F. Moens. Finding the Best Picture: Cross-Media Retrieval of Content. In *Proc. of ECIR 2008*, pages 539–546, 2008.

[8] M. Di Iulio and A. Messina. Use of Probabilistic Clusters Supports for Broadcast News Segmentation. In *DEXA Workshops*, pages 600–604, 2008.

[9] M. Farrús, P. Ejarque, A. Temko, and J. Hernando. Histogram equalization in svm multimodal person verification. In *Proc. of the Intl. Conf. on Advances in Biometrics*, pages 819–827, 2007.

[10] M. Henzinger, B. wei Chang, B. Milch, and S. Brin. Query-free news search. In *Proc. of WWW03*, pages 1–10, 2003.

[11] V. Kashyap and A. Sheth. Semantic and schematic similarities between database objects: a context-based approach. *The VLDB Journal*, 5(4):276–304, 1996.

[12] X. Li, J. Yan, Z. Deng, L. Ji, W. Fan, B. Zhang, and Z. Chen. A novel clustering-based RSS aggregator. In *Proc. of WWW07*, pages 1309–1310, 2007.

[13] A. Messina. An application of NLP and audiovisual content analysis for integration of multimodal databases of current events. In *Proc. of NLDB08*, pages 350–351, 2008.

[14] A. Messina, R. Borgotallo, G. Dimino, D. A. Gnota, and L. Boch. Ants: A complete system for automatic news programme annotation based on multimodal analysis. In *Proc. of WIAMIS 2008*, 2008.

[15] H. T. Pao, Y. Y. Xu, S. C. Chung, and H. C. Fu. Constructing and application of multimedia tv news archives. In *Intl. Workshop on Multimedia Content Analysis and Mining*, pages 151–160, 2007.

[16] C. Wang, M. Zhang, S. Ma, and L. Ru. Automatic online news issue construction in web environment. In *Proc. of WWW08*, pages 457–466, 2008.

[17] D. Webster, W. Huang, D. Mundy, and P. Warren. Context-orientated news filtering for web 2.0 and beyond. In *Proc. of WWW06*, pages 1001–1002, 2006.

[18] X. Wu, J. Li, Y. Zhang, S. Tang, and S. Y. Neo. Personalized multimedia web summarizer for tourist. In *Proc. of WWW08*, pages 1025–1026, 2008.

[19] C. Xu, J. Wang, H. Lu, and Y. Zhang. A novel framework for semantic annotation and personalized retrieval of sports video. *IEEE Trans. on Multimedia*, 10(3):421–436, 2008.

[20] Y. Zhang, C. Xu, Y. Rui, J. Wang, and H. Lu. Semantic event extraction from basketball games using multi-modal analysis. In *Proc. of the ICME 2007*, pages 2190–2193, 2007.