Web News Categorization using a Cross-Media Document Graph

João Magalhães[1], José Iria[2], Fabio Ciravegna[2]

[1]Dep. Engenharia Electrónica Telecomunicações e Computadores
Instituto Superior de Engenharia de Lisboa
Lisbon, Portugal

[2]Department of Computer Science
The University of Sheffield
Sheffield S1 4DP, UK

(jmag@deetc.isel.ipl.pt, j.iria@dcs.shef.ac.uk, fabio@dcs.shef.ac.uk)

ABSTRACT

In this paper we propose a multimedia categorization framework that is able to exploit information across different parts of a multimedia document (e.g., a Web page, a PDF, a Microsoft Office document). For example, a Web news page is composed by text describing some event (e.g., a car accident) and a picture containing additional information regarding the real extent of the event (e.g., how damaged the car is) or providing evidence corroborating the text part. The framework handles multimedia information by considering not only the document's text and images data but also the layout structure which determines how a given text block is related to a particular image. The novelties and contributions of the proposed framework are: (1) support of heterogeneous types of multimedia documents; (2) a document-graph representation method; and (3) the computation of cross-media correlations. Moreover, we applied the framework to the tasks of categorising Web news feed data, and our results show a significant improvement over a single-medium based framework.

Categories and Subject Descriptors

H.3.1 [Content Analysis and Indexing]: Abstracting methods. E.2 [Data Storage Representations]: Composite structures.

General Terms

Algorithms, Measurement, Experimentation.

Keywords

Web news categorization, cross-media correlations, document-graph, cross-media documents.

1. INTRODUCTION

In many real-world environments, the nature and complexity of multimedia is very diverse. In particular, multiple document formats abound, such as Microsoft Office's, HTML, OpenDocument and PDF. Contrary to plain text, these formats typically carry a mixture of textual content, metadata about the text (e.g. style information), images, tables and other media objects, e.g., Figure 1. However, to minimise processing complexity, retrieval systems tend to simply strip documents down to their core textual format – typically sets or sequences of words. While this may be enough for tasks such as text categorization [20], where state-of-the-art approaches can afford to model the document simply as a bag of words to produce useful results, generally such stripping down process throws away layout or relational information which would otherwise provide valuable features to a retrieval algorithm [4, 16, 21]. This is particularly true with modern multimedia documents, where valuable features can sometimes be found on the structure layout and on the different media composing the document. For example, health care and aeronautics engineering are two examples of real-world applications demanding search engines supporting multimedia documents such as patient diagnostic reports [10, 24] or jet-engine maintenance reports [7].

Figure 1. Example of a Web news article.

1.1 Proposed Framework

To support a truly cross-media categorization framework, two of the most important challenges are: (1) the data representation issue arising from the fact that different input document formats model documents differently, making it hard for analysis systems to have access to an integrated view, across formats, of the text and its related metadata and media objects; and (2) the layout structure issue, i.e., the problem of establishing a relation between text elements (sections, paragraphs, sentences, etc.) and

ACM XXX-X-XXXXX-XXX-X/XX/XXXX.

images - in the more difficult cases it is hard even to understand whether an image and a text element are related at all.

Thus, it is in this setting that we propose a novel cross-media document-graph categorization framework that directly tackles each of the aforementioned challenges and exploits the multiple sources of information contained in a document. The framework processes information in a series of steps, as follows:

- **Document-graph computation:** a specific parser converts a multimedia document into the generic document-graph representation designed to ensure that any type of data is sufficiently represented for the purposes of running the inference algorithms over the representation induced by the corpus.
- **Computation of cross-media correlations:** a simple, yet informative, structural analysis algorithm to detect correlations between the different media elements.
- **Inference:** given a set of training document-graphs a model is estimated for each category. In a later phase, the same model is used on new documents to infer their category.

This allows us to formulate the hypothesis of this paper: *the document-graph can improve multimedia categorization precision-recall, by preserving not just text and images, but also the cross-media correlations between text blocks and images.*

1.2 Contributions

In our view the main contributions of the proposed multimedia document-graph framework are:

1. **Experiments repeatability:** software[1] and test data are fully available.
2. **Document-graph:** a canonical document representation graph that integrates data coming from heterogeneous formats and media. This representation is designed such that it is able to accommodate, for every supported document format, enough information to allow an inference algorithm(s) to run.
3. **Cross-media correlations:** a novel method for detecting HTML cross-media associations and quantifying the level of image and text block correlation.
4. **Improved effectiveness**: experimental results on a Web news dataset, Figure 1, show that our cross-media approach, exploiting features from more than one media, yields significant improvement over the best results obtained by the corresponding single-medium tasks.
5. **Heterogeneous multimedia documents:** generic enough to be applicable to several domains and multimedia data types, and yet enable improvements over the traditional single-medium setting.

The rest of the paper is organized as follows: after discussing related work we introduce the multimedia categorization framework. In section 3 we propose the document-graph representation of multimedia documents and in section 4 we describe how the cross-media correlation nodes are computed. The multimedia categories inference algorithm is discussed in section 5 and section 6 presents the evaluation of the proposed framework.

[1] Available at: http://runes.sourceforge.net, http://aleph-ml.sourceforge.net, and http://t-rex.sourceforge.net.

2. RELATED WORK

Some research has been done to tackle the problem of extracting information from multimedia content, but very few cases address such heterogeneous content as we do in this paper. Literature on document processing discusses several approaches to extract structural information from PDF, HTML and other structured document types, see [15] for an overview.

The approaches proposed by Arasu and Garcia-Molina [1], Crescenzi et. al [6] and Rosenfeld et. al. [19] are based on templates that characterize each part of the document. These templates are either extracted manually or semi-automatically. Rosenfeld et. al. [19] devised a learning algorithm to extract information (author, title, date, etc) that relies on a general procedure for structural extraction. Their proposed technique enables the automatic extraction of entities from the document based on their visual characteristics and relative position in the document layout. They ignore text content and only use features such as fonts, physical positioning and other graphical characteristics to provide additional context to the information. Similarly, Fernandes et al. [12] describe a framework that computes the importance of each HTML page block for a given query. Their algorithm is based on the segmentation algorithm proposed by Yu et al. in [23] – an unsupervised web page segmentation algorithm called VIPS which constructs the content structure of a web page based on visual cues. Like the latter, our approach is based on a set of heuristics that extracts and preserves all structure information. Moreover, in contrast to these approaches, we implement an additional cross-media analysis aimed at discovering associations between images and paragraphs in the text.

The way documents are represented is another challenge in the problem that we address in this paper. Denoyer [8, 9] propose a statistical graph representation for the classification of structured documents. Their system classifies multimedia documents, where text and images are mixed together in HTML pages. The multimedia document model in their experiments demonstrated significant improvement over the traditional bag of words document model. Zhuang [25] present a method for cross-media retrieval, in which cross-media features are integrated with multimedia data via a cross-reference graph model so as to improve retrieval accuracy progressively by learning associations between objects present in the model. We note that these two works have improved on the work in [19] by identifying the relations between different media objects . But whereas these approaches need to be trained using supervised structure learning algorithms, our canonical representation of documents is completely unsupervised and generic enough to support a wide range of multimedia data types.

Another problem that we tackle in this paper is how to build a single classifier from the low-level features that come from the different single-medium parts of a multimedia document.

Previous approaches use co-training [3] or ensemble algorithms [14] that train different classifiers on single-medium feature vectors and combine the classifiers through a voting scheme to produce a single classifier with better accuracy. More recent approaches concatenate the single-medium feature vectors into a single cross-media vector, see Magalhães and Rüger [17]. Although we also concatenate single-medium features into a single feature vector, our approach differs from the previous ones because we construct the cross-media feature vector by also taking into account the confidence that a given text-image pair are associated, in the form of weights affecting the text tokens that participate in the identified relation.

The idea of using features from text and images has also been applied to tasks other than classification. For example, [2] use a generative hierarchical model for clustering image collections which integrates semantic information provided by associated text and visual information provided by image features. The data is modeled as being generated by a fixed hierarchy of nodes, with leafs of the hierarchy corresponding to clusters. The work in [5] combines textual and visual statistics in a single index vector for content-based search of a WWW image database. Textual statistics are captured in vector form using latent semantic indexing (LSI) based on text in the containing HTML document, while visual statistics are captured in vector form using color and orientation histograms. The authors show that the combined approach allows improved performance in conducting content-based search. In addition to text and image features, our work makes use of cross-media correlations.

Other classification tasks related to our work include classifying images with the help of text. For example, [11] develop an image annotation model on a dataset of pictures naturally embedded into news articles and show that using captions as a proxy for annotation keywords can remove the overhead of manual annotation, and also demonstrate that the news article associated with the picture can be used to boost image annotation performance. Our task is the inverse, that is, we classify documents with the help of images.

3. DOCUMENT GRAPH

Due to the advent of the WWW, the HTML format is probably the most common electronic resource available nowadays. An HTML document is often rich in media objects and layout information. Compared to plain text, HTML adds a layer of metadata mostly containing spatial (layout) information, on top of the text. HTML allows the creation of complex structures and relations, in particular relations between the different media elements within the document. In order to capture such complexity, we adopted the canonical graph representation for a document as depicted in Figure 2. The document-graph represents a multimedia document with a set of text nodes, image nodes and cross-media nodes. Formally, each document is defined as

$$d_n = \{T_n, I_n, X_n, E_n\}, \quad (1)$$

where its elements are:

- a set $T_n = [T_{n,1},...,T_{n,M}]$ of M text data nodes, where each node contains a meaningful text block (sentence, paragraph, title, heading, caption or alt_text) and the corresponding feature vectors describing the text content;
- a set $I_n = [I_{n,1},...,I_{n,N}]$ of N image data nodes, where each node contains an image of the document and the corresponding feature vectors describing the visual aspects of the image;
- a set $X_n = [X_{n,1},...,X_{n,O}]$ of O cross-media data nodes, relating text nodes and image nodes. This node type contains a correlation value indicating the probability that both referred nodes concern the same information; and
- a set E_n of edges between the different data nodes.

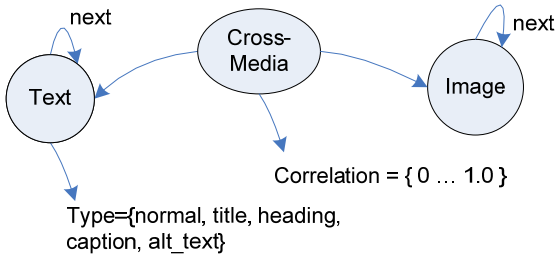

Figure 2. The graph based document representation.

3.1 An Example

Figure 3 illustrates the document-graph extracted from the HTML page shown in Figure 1. The graph is created by parsing the document's HTML code and converting it to the canonical representation.

First, all text nodes are created by extracting all sentences and creating a node in the graph for each sentence. The sequence of nodes on the left in the graph of Figure 3 depicts all nodes representing the text part of the document. The single node on the right represents the only picture in this document. Cross-media correlations are then computed by first analysing the layout of the document and only considering the text nodes in the neighbourhood of the image node, in this case nodes *Text2* to *Text6*. This creates cross-media nodes between these nodes and *Image1*.

Further processing computes the similarity between the image caption and the text of each text node to better estimate the degree of pairwise correlation between single-medium nodes. In the following we provide an intuitive description of the process, providing a more formal description in the next section. For example, *Text2* and *Image1* exhibit a high value of cross-media similarity because *Text2* is close (in the DOM tree sense) to the image and it contains words that are present both in the image caption and alt-text. Node *Text4* also has a high similarity value to the image because this sentence mentions "animal" and "scientist", which are also both present in the image caption and alt-text. The cross-media association between nodes *Text6* and *Image1* is directly extracted from the layout, so there is no need to compute a similarity.

Finally, cross-media nodes will then affect the referenced text nodes by giving more weight to the text present in those nodes. This will increase the importance of text blocks that reflect, or are related to, the content of an image in the document.

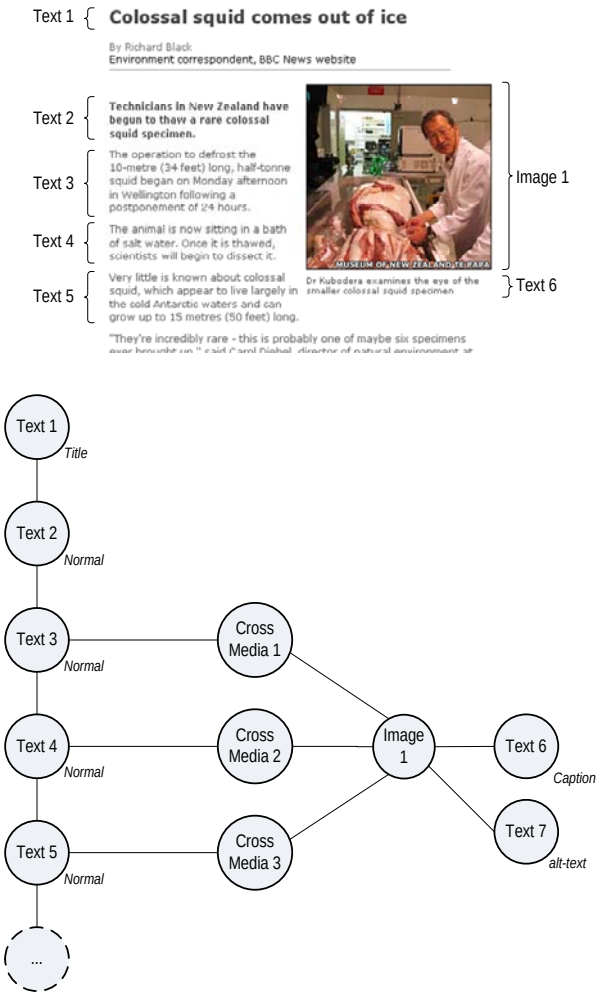

Figure 3. Example of a document parsing analysis and the corresponding document graph.

3.2 HTML Parsing

To create the document-graph, one first needs to parse the specific document format. Since most document formats (e.g., Microsoft Office and PDF) can be easily converted to the HTML format, we focus on describing support for the HTML format in this paper.

The implemented HTML processor preserves most of the original information when converting it into the canonical graph representation. However, not all information on a Web page is important for the task at hand, e.g., advertisements and navigational links. Thus, our HTML processing algorithm is based on a set of heuristics that filters out irrelevant content. For example, the Web page in Figure 3 is obtained from the web page in Figure 1 by discarding navigational links, adverts and other such information. The extraction of the relevant corpus starts by converting the document from HTML into a well-formed XHTML and parsing it with the SAX[2] parser. The following rules strip the document of unwanted content:

- **News corpus identification:** The algorithm parses the XML tree to locate the tree branch containing the news body. The news body is usually inside a *div* element with specific attributes. This removes the main parts outside the news body.
- **Noisy-structures removal:** Non-corpus content (e.g., videos, reader's comments) and irrelevant content (e.g., navigational links, adverts) are removed by locating elements corresponding to particular patterns known in advance.
- **Noisy-images removal:** Some images in the corpus are too small to be processed or are just stylistic images (e.g., an icon). Images with less than 200 pixels are ignored and images with a URLs pointing to a specific location (e.g., location where all formatting images are stored) are ignored as well.

This process generates a clean XHTML document that serves as the basis for the creation of the graph. Once the clean XHTML document is obtained, a second pass is done to extract text and image nodes as will be detailed in the next sections.

3.3 Text Nodes

Text nodes are generated by (1) analysing the layout structure of the HTML Web page and (2) parsing the news text body to extract sentences and (3) processing text data with standard text processing techniques. The details follow:

- **Formatting based analysis.** Style and layout information, such as *headings*, *bolds* and *divs*, define the structure of a document. Here, we use standard HTML formatting tags to guide the extraction of the news text, section titles (tags `<h1>`, `<h2>` and `<title>`), alt-text (`<img alt-text="…">`) and image captions (in our case, an image is always inside a `<div>` element together with a text element corresponding to the caption text). This creates text nodes corresponding to titles, captions and alt-text.
- **Text body analysis.** Textual cues like punctuation provide further information to segment the text into paragraphs and sentences. As a result, documents can be represented by text nodes capturing different levels of abstraction: one text node per document (useful for document classification), one node per document section and one node per paragraph (useful for entity recognition or relation extraction tasks). The sentence level of abstraction is fundamental to determine the cross-media nodes in the approach here presented, as will be detailed in section 3.5. This step creates text nodes corresponding to sentences in the news body.
- **Text data processing.** Once the text nodes have been created according to the previous heuristics, text data is processed with standard text processing techniques [22]: stop words and infrequent words are removed from the text corpus (to avoid over-fitting). After this, the Porter stemmer reduces words to their morphological root – the so-called "*terms*". The resulting text nodes contain a histogram of the terms, e.g, the vector

[2] HTML cleaning, SAX parsing and XML data processing are done with tools provided by the Apache Software foundation.

$$T_{n,s} = \left[t_1, ..., t_V \right], \quad (2)$$

represents node s of document n, where each component is the frequency of the corresponding term.

3.4 Image Nodes

As mentioned previously, all images larger than 200 pixels are considered relevant for the news at hand and indicative of the news category. For each `<img>` element, an image node is created in the document-graph. The contents of the image node i of the document n are represented as the feature vector

$$I_{n,i} = \left[I_{n,i}^{HSV}, I_{n,i}^{Gabor}, I_{n,i}^{Tamura} \right], \quad (3)$$

where each component corresponds to the following visual features:

- **Color features:** images are split into 9 equal tiles and an HSV histogram per colour channel with 256 bins is computed.
- **Gabor texture features:** are computed with a bank of filters in 8 directions and 6 scales for the entire image. We consider the mean and the variance of the output of each filter.
- **Tamura texture features:** images are split into 9 equal tiles and the three Tamura texture features are computed (contrast, directionality and coarseness).

For more details on the visual features please refer to [13].

3.5 Cross-Media Nodes

Cross-media nodes capture the relatedness between a given image and surrounding text. The implemented method uses a mixture of layout information and text-based similarity to create cross-media nodes. The next section details this process.

4. CROSS-MEDIA CORRELATIONS

A multimedia document can express an idea across different media, with each text paragraph and image offering support on different aspects of the overall exposed information. Document structure and cross-references (e.g. captions) can suggest how each text sentence relates to each image. For example, in Figure 3 the node *Text6* is the caption of node *Image1*, and that can be easily determined from the style of the document (both elements are inside the same `<div>`).

However, it is not always straightforward to know which elements refer to the same information. In this section, we detail how text-image associations (cross-media correlations) are inferred from the layout structure, and from the image's caption and alternative text.

4.1 Layout based Correlations

Taking advantage of the layout information is arguably the most obvious way of associating an image to a set of text paragraphs. The XML tree of the XHTML document provides proximity information that can be used to infer valid cross-media associations. The detection of layout-based cross-media associations is done on the XML tree by imposing a window over the text nodes, centred on the image node.

Formally, the layout distance between an image I_i and a text block T_s is defined as

$$f_{\mathrm{L}}\left(I_i, T_s\right) = 1 - \frac{\mathrm{NodeDist}\left(I_i, T_s\right)}{MaxWindowSize + 1}, \quad (4)$$

where $\mathrm{NodeDist}\left(I_i, T_s\right)$ is the number of nodes between an image node I_i and a text node T_s (for simplicity, we dropped the index n corresponding to the document), and $MaxWindowSize$ is the maximum window size covering the text nodes around the considered image node. To avoid missing the true cross-media relationships, in this work we imposed a window of length equal to the whole document – this way we compute the distance between an image and all sentences in the news.

4.2 Text based Correlations

In some cases, the layout similarity is not enough to relate two single-media nodes. Cross-references are another way of associating a sentence or a paragraph to an image. The following example illustrates this case:

*"The **squid** is now sitting in a bath of salt water…",*

where the picture is actually showing the mentioned squid. To detect these associations, techniques for measuring text similarity can determine the level of relatedness between a sentence T_k and an image i through its text caption $T_{i,c}$ and its alternative text $T_{i,a}$ (for simplicity, we dropped the index n corresponding to the document).

Formally, the correlation between a sentence and an image is measured as the cosine distance between the sentence T_k and the image's associated text $T_i = T_{i,c} + T_{i,a}$,

$$f_{\mathrm{T}}\left(T_i, T_s\right) = 1 - \frac{T_i}{\left\|T_i\right\|} \cdot \frac{T_s}{\left\|T_s\right\|}. \quad (5)$$

4.3 Total Correlation

The layout-based correlation might miss a correct correlation if the text node is too far from the image. To tackle this issue, the text-based correlation compensates the layout distance by giving more weight to relevant nodes. Hence, by merging the two methods for detecting cross-media relations between an image node I_i and a text node T_s we are able to quantify the degree of correlation between two single media elements.

The total cross-media correlation between two single media nodes is given by the function

$$\gamma_{i,s} = \frac{f_L\left(i_i, T_s\right) + f_{\mathrm{T}}\left(T_i, T_s\right)}{2}. \quad (6)$$

Formally, a cross-media node k is represented as

$$X_k = \left\{ I_i, \quad T_s, \quad \gamma_{i,s} \right\}, \quad (7)$$

relating nodes I_i and T_s, and quantifying their degree of correlation with the function $\gamma_{i,s}$.

5. WEB NEWS CATEGORIZATION

Our aim is now to infer the category of a given multimedia document d_n from its cross-media document-graph. To complete this task we follow a probabilistic approach, i.e.

$$p\left(w_l \mid d_n = \left\{T_n, I_n, X_n, E_n\right\}, \beta_l\right), \tag{8}$$

where β_l corresponds to the model of the news category w_l from the set

$$\mathcal{W} = \left\{w_1, ..., w_L\right\} \tag{9}$$

of L categories. In this setting, we define a collection

$$\mathcal{D} = \left\{d_1, d_2, \ ..., \ d_N\right\}, \tag{10}$$

of N multimedia documents split into a training set to learn the category models and a test set for evaluation. To simplify the exposition, we shall assume multimedia documents have only one image – the treatment is easily extended to the general case. In this probabilistic setting, a document d_n is equivalent to the vector

$$d_f^n = \left[\sum_{s=1,...,M} \gamma_{1,s} \cdot T_{n,s} \qquad I_{n,1}\right] \tag{11}$$

which includes all text content (the sum over all T nodes), an image feature vector, and the cross-media correlation $\gamma_{1,s}$ which weighs sentences according to their relevance to image $I_{n,1}$. The probabilistic framework of equation (8) was formally implemented as

$$\log\frac{p\left(w_l \mid d_f^n\right)}{p\left(\overline{w_l} \mid d_f^n\right)} = \log\frac{p\left(w_l\right)}{p\left(\overline{w_l}\right)} + \sum_i p\left(d_{f,i}^n\right)\beta_{l,i} \tag{12}$$

where $\overline{w_l}$ indicates the non-presence of the category w_l, $d_{f,i}^n$ is the i^{th} dimension of the document vector d_f^n, and $\beta_{l,i}$ is the i^{th} dimension of the category linear model. Furthermore, the model of a category is computed as

$$\beta_{l,i} = \log\frac{E\left[d_{f,i} \mid w_l\right]}{E\left[d_{f,i} \mid \overline{w_l}\right]}. \tag{13}$$

The interpretation of this equation is straightforward: the dimension $\beta_{l,i}$ is close to zero if the i^{th} dimension of $d_{f,i}$ is irrelevant for the category, positive if it is frequent, and negative if it is rare. This way, when evaluating unseen samples each dimension will have a low or high contribution to the detection of the category.

Finally, to recover the more general case where documents have more than one image, we simply average the output of $p\left(w_l \mid d_f^n, \beta_l\right)$ for each image in the document.

6. EXPERIMENTS

To evaluate our framework, we conducted a categorization experiment on BBC Web news articles that were obtained via a RSS feed. Results were assessed in a retrieval scenario.

6.1 BBC News Data[3]

The experiment uses BBC Web news articles that were obtained between the 2nd of May 2008 and the 4th of June 2008. The category of each news article is obtained via the news category assigned by BBC journalists. On the BBC website, news URLs are organised according to category, thus, it is possible to extract the category from the article's URL. There are a total of 44 categories, listed in Table 1.

Table 1. BBC news data categories.

/uk	/sport
/uk_politics	/sport/football/eng_div_1
/magazine	/sport/cricket
/health	/sport/cricket/england
/technology	/sport/rugby_league
/business	/sport/football
/education	/sport/football/eng_prem
/entertainment	/sport/football/europe
/science	/sport/football/internationals
/science/technology	/sport/olympics
/northern_ireland	/sport/rugby_union
/england	/sport/tennis
/england/london	/sport/athletics
/england/manchester	/sport/cricket/counties
/scotland	/sport/motorsport
/wales	/sport/motorsport/formula_one
/world	/sport/motorsport/motorbikes
/world/americas	/sport/boxing
/world/middle_east	/sport/other_sports
/world/europe	/sport/other_sports/cycling
/world/south_asia	/sport/other_sports/horse_racing
/world/asia-pacific	
/world/africa	

We collected a total of 6,727 news articles and randomly split them into a set of 4,485 training documents and 2,242 test documents. Each news article belongs to just one category and most articles have at least one image. It is worth noting that this dataset is different from other news datasets such as Reuters-21578. The latter contains pure text documents, in contrast with the BBC Web news articles, which are multimedia documents with images and structure.

6.2 Experiment Setup

Documents are first transformed into the document-graph and their cross-media correlations are computed according to equation (6). Category models are learned from the training documents vectors, equation (11), and computed according to equation (13). Once the system is trained, we followed traditional document retrieval by category evaluation: for each category we ranked the test documents according to equation (12) and evaluated the rank with precision-recall curves, average precision, precision after 10 retrieved documents and precision after 30 retrieved documents. The means across all queries are computed from the results per query: mean average precision,

[3] Available after publication.

mean precision at 10 and mean precision at 30. This procedure was repeated for (i) only the text data of documents, (ii) only images data, and (iii) text, images and cross-media data.

6.3 Results and Discussion

Table 2 presents the multimedia retrieval by category comparison results for the three settings: text-only, images-only and cross-media. Note that cross-media results are always better in all three measures, showing how important it is to consider the relations between different modalities. Figure 4 presents the same results in a bar chart for better comparison.

Table 2. Retrieval results for different modalities and cross-media.

	MP@10	MP@30	MAP
Images	7.27 %	5.70 %	4.68 %
Text	70.90 %	51.07 %	61.66 %
Cross-Media	73.18 %	52.02 %	62.30 %

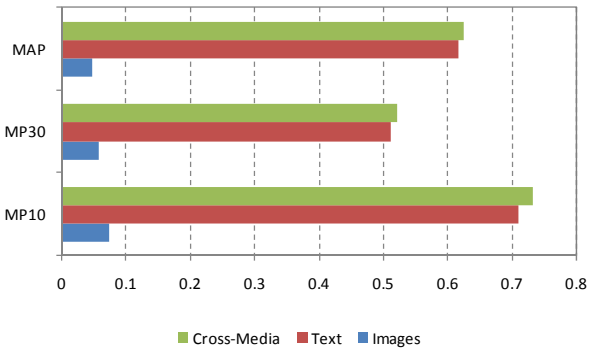

Figure 4. Retrieval results for different single-medium and cross-media.

Not surprisingly, image results are always much lower than any other settings. This observation is justified by the fact that some categories cannot be actually discriminated from just images. For example, there is virtually little difference among the pictures of the categories "/England" "/England/London", "/Scotland" "/Northern_Ireland", "/England/Manchester", and "/Wales". Thus, images only contain information to discriminate between categories like "/sports" and "/uk_politics", i.e., broader categories.

The precision-recall graph on Figure 5 provides another view on how the cross-media document-graph performs. It is interesting to see that cross-media is much better at the beginning of the rank (recall < 10%), which is where most users look at. Also, at the mid-range of the rank one sees that cross-media is clearly better than just text. Moreover, it is actually exciting to see how image data contribute to the cross-media results despite the fact that image-only results are so low when compared to the other results. We believe that this strengthens our hypothesis and gives evidence that cross-media correlation is an important aspect of multimedia documents.

Finally, it must be noted that categories are hierarchical and some news may be placed in one category but in fact they can belong to more than one category. This introduces noise in the training phase and in the testing phase, and implies that results would improve if we clean these incomplete hierarchical category annotations.

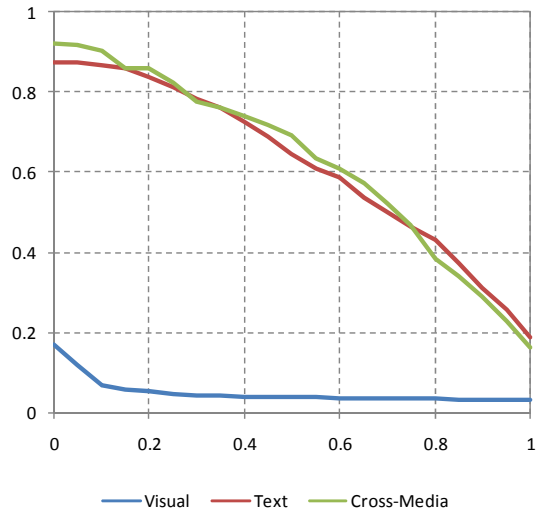

Figure 5. Precision-recall graphs for different modalities and cross-media.

7. CONCLUSIONS AND FUTURE WORK

We have presented a cross-media information extraction framework that is able to exploit information about the same concept across different media and contained in different parts of a document (e.g., a text paragraph or an image). The framework handles multimedia data by using features from text, images and layout simultaneously, and in a novel way.

As discussed in the introduction, the contributions of this paper are the document-graph, the cross-media correlations, the improved effectiveness and the support of heterogeneous multimedia documents. Moreover, we provide all the software and data. The experimental results obtained by applying the framework to the tasks of categorising Web news feed data validate our hypothesis: *the document-graph can improve multimedia categorization precision-recall by preserving not just text and images of documents but also the cross-media correlations between text blocks and images*. It also validated the framework's support of Web content, the construction of the document-graph, and the computation of the cross-media correlations. More importantly it shows that exploiting the correlations between features of different media yields a significant improvement over the best results obtained by the corresponding single-medium tasks.

The presented results lead us to several research questions that need to be further investigated and studied. For example: (i) we plan to improve the cross-media correlations by embedding more complex text processing techniques such as named entities, entities' relations, and more meaningful features [18]; (ii) study in more detail how the contribution of features coming from the different media is affected by the amount of training data available; (iii) develop an inference algorithm that explicitly

considers more than one image and the relations between the images; and (iv) study how implicit cross-references can be better detected, e.g., *"The squid, shown in **above picture**, is now sitting in a bath of salt water …"* - in this example the "*above picture*" text tells us that the current sentence is related to the image.

8. ACKNOWLEDGEMENTS

This research was partially funded by the X-Media project (www.x-media-project.org) sponsored by the European Commission as part of the Information Society Technologies (ITS) programmer under EC grant number ITS-FPO6-026978.

9. REFERENCES

[1] A. Arasu and A. H. Garcia-Molina, "Extracting structured data from Web pages " in *ACM SIGMOD conf. on management of data* San Diego, California 2003.

[2] K. Barnard and D. A. Forsyth, "Learning the semantics of words and pictures," in *Int'l Conf. on Computer Vision*. vol. 2 Vancouver, Canada, 2001, pp. 408-415.

[3] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Computational Learning Theory* Madison, WI, USA, 1998.

[4] T. M. Breuel, "Information extraction from HTML document by structural matching," in *Int'l Workshop on Web Document Analysis* Edinburgh, UK, 2003, pp. 11–14.

[5] M. L. Cascia, S. Sethi, and S. Sclaroff, "Combining Textual and Visual Cues for Content-based Image Retrieval on the World Wide Web," in *IEEE Workshop on Content-based Access of Image and Video Libraries with the IEEE Conf. on Vision and Pattern Recognition* Santa Barbara, California, 1998.

[6] V. Crescenzi, G. Mecca, and P. Merialdo, "RoadRunner: towards automatic data extraction from large Web sites," in *Int'l Conference on Very Large Data Bases*, 2001.

[7] A.-S. Dadzie, R. Bhagdev, A. Chakravarthy, S. Chapman, J. Iria, V. Lanfranchi, J. Magalhães, D. Petrelli, and F. Ciravegna, "Applying Semantic Web technologies to knowledge sharing in aerospace engineering," *Journal of Industrial Manufacturing*.

[8] L. Denoyer and P. Gallinari, "Bayesian network model for semi-structured document classification," *Information Processing and Management*, vol. 40, pp. 807–827, June 2004.

[9] L. Denoyer, P. Gallinari, J.-N. Vittaut, S. Brunesseaux, and S. Brunesseaux, "Structured multimedia document classification," in *ACM DOCENG* Grenoble, France, 2003.

[10] S. Ebadollahi, L. Xie, S.-F. Chang, and J. R. Smith, "Visual Event Detection Using Multi-Dimensional Concept Dynamics," in *IEEE International Conference on Multimedia and Expo* Toronto, Canada, 2006.

[11] Y. Feng and M. Lapata, "Automatic Image Annotation Using Auxiliary Text Information," in *ACL HLT* Columbus, Ohio, USA, 2008.

[12] A. Haubold and A. Natsev, "Web-based information content and its application to concept-based video retrieval," in *ACM Conf. on Image and Video Retrieval* Niagara Falls, Canada, 2008.

[13] P. Howarth and S. Rüger, "Evaluation of texture features for content-based image retrieval," in *Int'l Conf. on Image and Video Retrieval* Dublin, Ireland, 2004, pp. 326-324.

[14] D. Joshi, M. Naphade, and A. Natsev, "Semantics reinforcement and fusion learning for multimedia streams," in *ACM international conference on Image and video retrieval* Amsterdam, The Netherlands, 2007.

[15] A. H. F. Laender, B. A. Ribeiro-Neto, A. S. d. Silva, and J. S. Teixeira, "A brief survey of Web data extraction tools," *ACM SIGMOD Record*, vol. 31 pp. 84-93.

[16] G. Maderlechner and P. Suda, "Information extraction from document images using white space and graphics analysis," in *Joint IAPR Int'l Workshop on Advances in Pattern Recognition*, 1998, pp. 468-474.

[17] J. Magalhães and S. Rüger, "Information-theoretic semantic multimedia indexing," in *ACM Conf. on Image and Video Retrieval* Amsterdam, The Netherlands, 2007.

[18] C. Manning and H. Schütze, *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press, May 1999.

[19] B. Rosenfeld, R. Feldman, and J. Aumann, "Structural extraction from visual layout of documents," in *ACM Conf. on CIKM* McLean, Virginia, USA 2002.

[20] F. Sebastiani, "Machine learning in automated text categorization," *ACM Computing Surveys*, vol. 34, pp. 1-47.

[21] C. Shin and D. Doermann, "Classification of document page images based on visual similarity on layout structures," in *SPIE Vol. 3967, Document Recognition and Retrieval VII* San Jose, California, 2000, pp. 182-190.

[22] Y. Yang, "An evaluation of statistical approaches to text categorization," *Information Retrieval*, pp. 69-90.

[23] S. Yu, D. Cai, J.-R. Wen, and W.-Y. Ma, "Improving Pseudo-Relevance Feedback in Web Information Retrieval Using Web Page Segmentation," in *World Wide Web* Budapest, Hungary, 2003.

[24] X. S. Zhou, S. Zillner, M. Moeller, M. Sintek, Y. Zhan, A. Krishnan, and A. Gupta, "Semantics and CBIR: a medical imaging perspective," in *ACM Conf. on Image and Video Retrieval* Niagara Falls, Canada, 2008.

[25] Y. Zhuang, H. Shan, and F. Wu, "An approach for cross-media retrieval with cross-reference graph and PageRank," in *International Conference on Multi-Media Modelling* Beijing, China, 2006