Exploiting Cross-Media Correlations in the Categorization of Multimedia Web Documents

José Iria
Department of Computer Science
The University of Sheffield
j.iria@dcs.shef.ac.uk

João Magalhães
Department of Computer Science
Instituto Superior de Engenharia de Lisboa
jmag@deetc.isel.ipl.pt

Abstract

Valuable features characterizing a document's content can often be found in the structure, the layout and the relationships between the different media comprising the document. In this paper, we present a novel approach to exploiting cross-media correlations from multimedia documents, and validate it in the context of a classification task over multimedia web news documents. The experimental results empirically show that, by preserving not just text and images but also the cross-media correlations between text elements and images, we can improve accuracy with respect to traditional approaches that ignore cross-media correlations. An advantage of our approach is that it makes almost no assumptions about the way multimedia content is modeled, and is thus widely applicable.

1 Introduction

The nature and complexity of modern multimedia documents presents both challenges and opportunities for automated processing. Contrary to plain text, multimedia documents typically carry a mixture of text, images, tables and metadata about the content (e.g., style information). Valuable features characterizing the document content can often be found in the structure, the layout and the relationships between the different media comprising the document. However, mining systems tend to simply strip documents down to their core single-medium formats, e.g. treating the document as a set or sequence of words plus a separate set or sequence of images, throwing away valuable cross-media features that would otherwise be valuable to the mining algorithm [Maderlechner and Suda, 1998; Breuel, 2003]. In fact, these traditional simplifying assumptions about document content pre-processing are not adequate for more demanding real-world applications such as information retrieval from patient diagnostic reports [Ebadollahi et al., 2006; Zhou et al., 2008] or jet-engine maintenance reports [Dadzie et al., 2008].

Instead of treating each medium separately, we would like to exploit correlations *across the media* elements. However, this is problematic for two main reasons. First, modern multimedia repositories consist of collections of content modeled in a variety of ways, since several document formats (e.g., HTML, Microsoft Office's, OpenDocument, PDF) have become ubiquitous. This complicates the development of multimedia analysis methods due to the lack of an integrated view, across document formats, of the several media elements in the corpus. More importantly, to exploit cross-media correlations there is also the problem of how to determine the relationship between those media objects. For example, typically in news articles one or more images are presented that illustrate the news story, and features such as their caption or their disposition with respect to the text flow are not random, but rather choices made by the editor according to conventions defined by professional and/or web communities, whether explicitly or implicitly.

In this paper, we present a novel approach to exploiting cross-media correlations from multimedia documents with the goal of improving mining systems' accuracy in classifying multimedia documents. Our approach only makes weak assumptions about the way multimedia content is expected to be modeled. Our hypothesis is that such minimal set of assumptions suffices to improve multimedia document classification. To test our hypothesis, we conduct a document classification experiment over a large collection of news stories collected from the Web, and empirically show an increase in accuracy when comparing our approach against approaches that do not exploit cross-media correlations.

To support our method, we have developed a framework that processes the content in a series of steps, as follows:

- **Document-graph representation**: format-specific parsers convert a multimedia document into a canonical 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. The learned model may then be used on new documents to infer their category.

In our view, the main contributions of the proposed approach are fourfold. Firstly, we present a novel method for detecting cross-media associations and quantifying the

level of image and text elements' correlation in the context of multimedia document classification. Secondly, we propose a set of minimal assumptions about the way multimedia documents are modeled, resulting in a canonical document representation graph that integrates data coming from heterogeneous formats and media, and accommodates enough information to allow inference algorithms to exploit the cross-media correlations. Thirdly, experimental results show that our cross-media approach yields significant improvement over the best results obtained by the corresponding single-medium tasks. Finally, for repeatability, software[1] and test data are fully available.

The rest of the paper is organized as follows: in the next section we detail the proposed document-graph representation of multimedia documents; then we describe how cross-media correlations are computed; in section 4 the inference algorithm used for categorization is discussed, while section 5 presents and discusses the results of evaluating the approach; in section 6 we contrast our approach against related work; and finally we conclude with an outline of future work.

2 Document Graph

The document-graph represents a multimedia document using text nodes, image nodes and cross-media edges. Formally, each document is defined as

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

where its elements are:

- a set $T_n = [T_{n,1},...,T_{n,M}]$ of M (non-nested) text data nodes, where each node contains a meaningful text block and the corresponding feature vectors derived from its content;
- a set $I_n = [I_{n,1},...,I_{n,N}]$ of N image data nodes, where each node contains an image and the corresponding feature vectors derived from its visual characteristics; and
- a function $X_n : I \times T \rightarrow \mathbb{R}$ of O cross-media edges, where each edge quantifies the relation between a text node and an image node. They contain a correlation value that expresses the likelihood that both referred nodes concern the same information.

This data representation model captures the essential information to perform cross-media classification and is independent of the way multimedia documents are modeled when stored in their original formats.

2.1 Multimedia Document Processing

The first step in the creation of the document-graph consists in parsing the documents. Although we have developed parsers for several formats, namely OpenDocument, PDF and HTML, here we discuss processing of the latter format only, noting that, for our purposes, the principles are identical regardless of format.

We use a set of heuristics to perform web content cleaning [Li and Ezeife, 2006][Hofmann and Weerkamp, 2007], that is, to filter out irrelevant content from the web pages, e.g., advertisements and navigational links in Web pages. The extraction of the "relevant corpus" starts by converting the document from HTML into a well-formed XHTML. The following rules strip the document of unwanted content:

- **Main body identification**: The pre-processor parses the XML tree to locate the tree branch containing the main body of the content. This involves handcrafting simple corpus-specific patterns over the XML tree.
- **Noisy-structures removal**: Content such as videos, comments, navigational links, adverts, and so on, remaining in the main body section, is removed, again using corpus-specific patterns.
- **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 URL 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 document-graph. Once the clean document is obtained, a second pass is done to extract text and image nodes, as will be detailed in the next sections.

2.2 Text Nodes

Text nodes are generated by (1) analysing the layout structure of the document and (2) parsing the textual content to extract sentences and (3) processing text data with standard text processing techniques. The details follow:

- **Formatting-based analysis**. Style and layout information define the structure of a document. In the case of XHTML, we use standard formatting tags to guide the extraction of the news text, section titles (tags <*h1*>, <*h2*> and <*title*>), alternative text for an image (<*img alt-text="..."*>) and image captions (via corpus-specific patterns). This creates the text nodes for titles, captions and "alt-text".
- **Text body analysis**. Textual cues like punctuation provide further information to segment the text. This step creates text nodes corresponding to sentences in the document, keeping the information about their sequence (in the form of *next-sentence* edges).
- **Text processing**. Standard text processing techniques [Yang, 1997] are applied: stop words and infrequent words are removed from the text corpus, to avoid over-fitting. After this, we apply the Porter stemmer to reduce words to their morphological root – the so-called "terms". The resulting text nodes contain a histogram of the terms, i.e., the vector

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

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

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

2.3 Images Nodes

For each <*img*> element, an image node is created in the document-graph. Information about the image sequence is kept in the form of *next-image* edges between image nodes, and information about immediately adjacent text nodes in the form of *previous*- and *next-sentence* edges (for XHTML this is done via the analysis of the DOM tree).

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 [Howarth and Rüger, 2004].

2.4 Cross-Media Edges

Cross-media edges store the likelihood that images and text nodes in the graph concern the same information. The next section describes the method employed to compute it.

3 Cross-Media Correlations

When a message or idea is conveyed through the different media in a multimedia document, each text paragraph and image offers support to different parts or aspects of the message. Because a full text typically expresses several ideas, the problem resides in trying to understand which media elements refer to the same information. In this work, we propose a method to do this in an unsupervised way and making very general assumptions about the document model, namely: that it is possible to obtain the sequence of text and images, and that images may have associated text like captions.

3.1 Layout-based Correlations

The sequence information in the document graph provides proximity information that can be used to infer layout-based cross-media associations. This is achieved by imposing a window over the text nodes centered on a given 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. In the experiments reported in section 5, we set the window length to be equal to the document size, thereby computing the correlation between all sentences in the document and the given image.

3.2 Text-based Correlations

Cross-reference information is another way of establishing the relation between a sentence and an image. They complement layout-based correlations, capturing those sentences that refer to an image not placed nearby (in terms of the document layout). To detect these associations, techniques for measuring text similarity can determine the level of relatedness between a sentence T_k and an image I_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 $T_i = T_{i,c} + T_{i,a}$, given by

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

3.3 Total Correlation

By merging the two methods above for detecting cross-media associations between an image node I_i and a text node T_s, we are able to quantify the degree of correlation between elements in the two media. Formally, the weight of a cross-media edge in the document-graph is given by the average of the two quantities defined above:

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

4 Multimedia Document Categorization

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

$$p\left(w_l \mid d_n = \left\{T_n, I_n, X_n\right\}, \beta_l\right), \quad (7)$$

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

$$\mathcal{W} = \left\{w_1, ..., w_L\right\} \quad (8)$$

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

$$\mathcal{D} = \left\{d_1, d_2, \ldots, d_N\right\}, \quad (9)$$

of N multimedia documents, split into a training set in order to learn the category models, and a test set for evaluation. To simplify the exposition, we shall assume

multimedia documents have only one image and extend the treatment to the general case *a posteriori*. In this probabilistic setting, a document d_n is equivalent to the vector

$$d_f^n = \left[\sum_{s=1,\ldots,M} \gamma_{1,s} \cdot T_{n,s} \qquad I_{n,1} \right] \tag{10}$$

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 (7) was formally implemented as

$$\log \frac{p\left(w_l \mid d_f^n\right)}{p\left(\overline{w_l} \mid d_f^n\right)} = \log \frac{p\left(w_l\right)}{p\left(\overline{w_l}\right)} + \sum_i p\left(d_{f,i}^n\right) \beta_{l,i} \tag{11}$$

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

$$\beta_{l,i} = \log \frac{\mathrm{E}\left[d_{f,i} \mid w_l\right]}{\mathrm{E}\left[d_{f,i} \mid \overline{w_l}\right]}. \tag{12}$$

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.

5 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.

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. A complete list of the categories can be obtained by downloading the dataset[2].

We collected a total of 6,727 news articles and randomly split them into 10 random sets of 4,485 training documents and 2,242 test documents for cross-validation. 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 plain text documents only, while the BBC Web news dataset used here consists of multimedia documents with images and structure.

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 (10), and computed according to equation (12). Once the system is trained, we followed traditional document retrieval by category evaluation: for each category we ranked the test documents according to equation (11) 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, mean precision at 10 and mean precision at 30. This procedure was repeated for (i) only text data, (ii) only image data, (iii) a simple concatenation of text and image features (multimedia), and (iv) text, image and cross-media correlation features.

Table 1 presents the results comparing multimedia retrieval by category in the four settings. Taking into account the cross-media correlations yields better results with respect to the other configurations, for all the three measures considered. Figure 1 presents the same results in a bar chart for better comparison.

Table 1. Retrieval results for the individual media, multimedia and cross-media configurations.

	MP@10	MP@30	MAP
Images	7.27 %	5.70 %	4.68 %
Text	70.90 %	51.07 %	61.66 %
Multimedia	70.23%	51.22%	61.68%
Cross-Media	73.18 %	52.02 %	62.30 %

Not surprisingly, image results are always much lower than the 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.

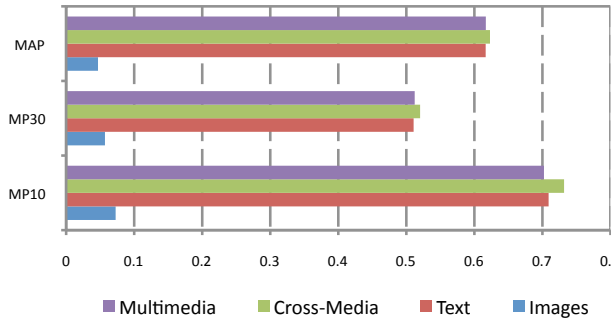


Figure 1. Retrieval results for the individual media, multimedia and cross-media configurations.

[2] http://www.dcs.shef.ac.uk/~jiria/datasets

Another interesting observation is that the simple concatenation of text and image features generally does not perform much better than text alone, and can actually result in lower precision, as is the case with measuring precision at ten. This does not happen with the proposed cross-media correlations method, which can be interpreted as a feature-weighing scheme that boosts the importance of text related to the image for learning the models.

The precision-recall graph on Figure 2 provides another view on how the cross-media approach 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, it can be seen that cross-media is better than just text and the concatenation of text and images. Moreover, it is very encouraging 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.

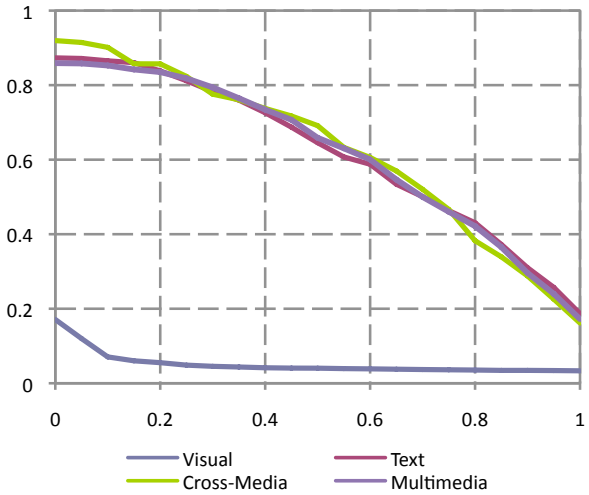


Figure 2. Precision-recall graphs for the individual media, multimedia and cross-media configurations.

6 Related Work

Several approaches to extract structural information from PDF, HTML and other structured multimedia document types can be found in the literature, see [Laender et al., 2002] for an overview. The approaches by [Arasu and Garcia-Molina, 2003], [Crescenzi et al., 2001] and [Rosenfeld et al., 2002] are based on templates that characterize each part of the document. These templates are either extracted manually or semi-automatically. Rosenfeld et al. 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. Like the latter, our approach is based on a set of heuristics that extracts and preserves all structure information. But, in contrast to these approaches, we implement an additional cross-media analysis step aimed at discovering associations between images and paragraphs in the text.

Another problem that we tackle in this paper is how to build a single classifier from the low-level features originating from the different single-medium elements of the multimedia document. Previous approaches use co-training [Blum and Mitchell, 1998] or ensemble algorithms [Joshi et al., 2007] 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, 2007]. 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, [Barnard and Forsyth, 2001] 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 [Cascia et al., 1998] 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, [Feng and Lapata] 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.

7 Conclusions and Future Work

We have presented a novel approach to exploiting cross-media correlations from multimedia documents and validated it in the context of a classification task over

multimedia web news documents. The experimental results empirically show that, by preserving not just text and images but also the cross-media correlations between text elements and images, the approach is able to improve the system performance, when compared against traditional approaches that ignore cross-media correlations. An advantage of our approach is that it makes almost no assumptions about the way multimedia content is modeled, and is thus widely applicable.

The presented results lead us to other research questions that need to be further investigated and studied. We plan to improve the cross-media correlations by embedding more complex text processing techniques such as named entities, entities' relations, and additional features. We will also study in more detail how the contribution of features coming from the different media is affected by the amount of training data available. Finally, we intend to develop an inference algorithm that explicitly considers more than one image, and the relations between the images.

References

[Arasu and Garcia-Molina, 2003] A. Arasu and A. H. Garcia-Molina. Extracting structured data from Web pages. *ACM SIGMOD Conf. on Management of Data*, San Diego, California 2003.

[Barnard and Forsyth, 2001] K. Barnard and D. A. Forsyth. Learning the semantics of words and pictures. In *Int'l Conf. on Computer Vision*, Vancouver, Canada, 2001.

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

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

[Cascia et al., 1998] 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*, Santa Barbara, California, 1998.

[Crescenzi et al., 2001] 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.

[Dadzie et al., 2008] 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 Intelligent Manufacturing,* Special issue on Knowledge Discovery and Management in Engineering Design and Manufacturing, 2008.

[Ebadollahi et al., 2006] 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 (ICME'06)*, Toronto, Canada, 2006.

[Feng and Lapata, 2008] Y. Feng and M. Lapata. Automatic Image Annotation Using Auxiliary Text Information. In *Proceedings of ACL 2008*, Columbus, Ohio, USA, 2008.

[Hofmann and Weerkamp, 2007] K. Hofmann and W. Weerkamp. Web Corpus Cleaning using Content and Structure. In *Building and Exploring Web Corpora*, Cahiers du Central (4), pages 145-154, UCL Presses, 2007.

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

[Joshi et al., 2007] D. Joshi, M. Naphade, and A. Natsev. Semantics reinforcement and fusion learning for multimedia streams. In *ACM Conf. on Image and Video Retrieval*, Amsterdam, The Netherlands, 2007.

[Laender et al., 2002] 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 Records*, vol. 31 pp. 84-93, 2002.

[Li and Ezeife, 2006] J. Li and C. Ezeife: Cleaning Web Pages for Effective Web Content Mining. In *17th International Conference on Database and Expert Systems Applications*, 2006.

[Maderlechner and Suda, 1998] 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*, pp. 468-474, 1998.

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

[Rosenfeld et al., 2002] B. Rosenfeld, R. Feldman, and J. Aumann. Structural extraction from visual layout of documents. In *ACM Conf. on Information Knowledge Management*, McLean, Virginia, USA 2002.

[Yang, 1997] Y. Yang. An evaluation of statistical approaches to text categorization. *Journal of Information Retrieval*, 1997.

[Zhou et al., 2008] 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.