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

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

João Magalhães <jmag@deetc.isel.ipl.pt>
Instituto Superior de Engenharia de Lisboa, Portugal

Organisations, Information and Knowledge

Motivation

- Multimedia docs carry a mixture of text, images, tables and metadata (e.g., style information)
- Mining systems tend to treat each medium separately
- There is potential to do better
  - Patient records [Ebadollahi et al., 2006; Zhou et al., 2008]
  - Jet engine maintenance reports [Dadzie et al., 2008]



- Valuable features in
  - the structure
  - the layout
  - the relationships between the different media
- We would like to exploit correlations *across the media* elements

Organisations, Information and Knowledge

Outline

- Challenges
- Related Work
- Proposed Approach
  - Document Representation
  - Text Analysis
  - Image Analysis
  - Cross-Media Analysis
  - Inference
- Experiments
  - Dataset
  - Methodology
  - Results
- Contributions
- Conclusions

Organisations, Information and Knowledge

Challenges

- How to take advantage of structural and layout info?
  - e.g., the position of an image with respect to the text flow
- What kind of relationships are there between the media objects in a document?
  - e.g., similarity of an image caption to the text
- Several multimedia formats have become ubiquitous (HTML, ODF, PDF, DOC, etc.)
  - Content modeled in a variety of way
  - Need to make simplifying assumptions that work regardless of underlying model

Related Work

- Multimedia document parsing
  - Template-based approaches [Arasu and Garcia-Molina, 2003], [Crescenzi et al., 2001] and [Rosenfeld et al., 2002]
    - Template characterizes each part of the document
    - Either manually built or extracted semi-automatically
  - Importance of HTML block to a given query [Haubold and Natsev, 2008]
  - Unsupervised web page segmentation algorithm [Yu et al., 2003]
  - We additionally perform cross-media analysis
- Learning from multiple feature types
  - Co-training [Blum and Mitchell, 1998]
  - Ensemble algorithms [Joshi et al., 2007]
  - Cross-media feature vectors using information theory-motivated clustering [Magalhaes and Ruger, 2007]
  - Our method can be seen as feature selection on text features
    - weights given by cross-media analysis

Related Work (2)

- Various other related tasks
- Classifying images with the help of text [Feng and Lapata, 2008]
  - Show that considering news article text where the image is contained can improve image annotation results
- Clustering image collections [Barnard and Forsyth, 2001]
  - generative hierarchical model using text and image features
- Supporting content-based search in a WWW image database [Cascia et al., 1998]
  - combines textual and visual statistics in a single index vector
  - Text features using LSI over HTML page text
  - Image features: color and orientation histograms

Proposed Approach

- **Goal:**
  - improving accuracy in classifying multimedia documents
- **Canonical representation**:
  - holds integrated representation (several formats, links, etc.) of documents
- **Computation of cross-media correlation**s
  - simple, yet informative, cross-media analysis method
- **Inference:**
  - estimate a model for each category

Organisations, Information and Knowledge

Contributions

- Canonical document representation encodes set of minimal assumptions about the way multimedia documents are modeled
  - integrates data coming from heterogeneous formats and media
  - accommodates enough information to allow inference algorithms to exploit the cross-media correlations
- Novel way of using the relation between image and text in a multimedia document
  - approach can be seen as a feature selection technique
- Empirical results showing small but consistent improvement over the best results obtained by the corresponding single-medium tasks
- Software and test data fully available. Code available from the following projects at Sourceforge:
  - http://runes.sourceforge.net
  - http://aleph-ml.sourceforge.net
  - http://t-rex.sourceforge.net

Organisations, Information and Knowledge

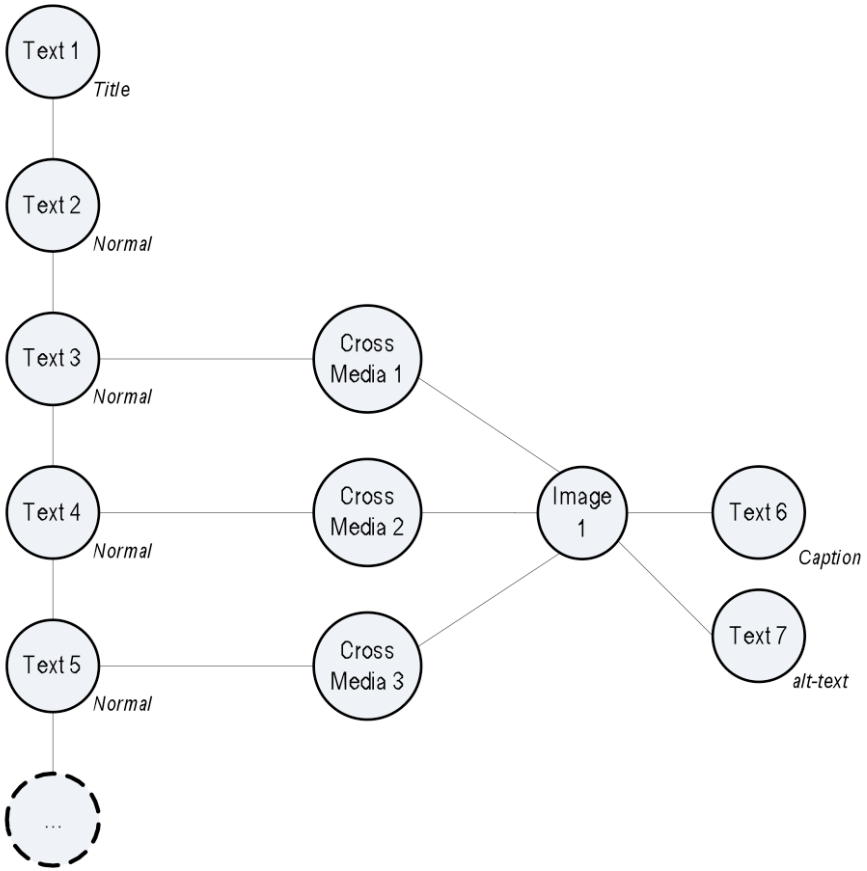
Canonical Representation

- Document

$$d_n = \{T_n, I_n, X_n\}$$

- Modeling:
  - Image nodes
  - Text nodes (several types)
  - Edge weights measure strength of relationship between text nodes and an image node





Text Analysis

- Analyse the metadata in the document content
  - format-specific processing
  - e.g., for HTML
    - section titles (tags *<h1>*, *<h2>* and *<title>*)
    - alternative text for an image (*<img alt-text="…">*)
    - image captions (via corpus-specific patterns)
- Parse the news text body to extract sentences
  - use textual cues such as punctuation
- Process text data with standard text processing techniques
  - stop words and infrequent words removed
  - Porter stemmer to reduce words to morphological root
- Create the various node types in the doc representation

Organisations, Information and Knowledge

Image Analysis

- Compute three types of visual features:
  - **Color features**
    - split images into 9 equal tiles
    - HSV histogram per colour channel with 256 bins
  - **Gabor texture features**
    - bank of filters in 8 directions and 6 scales for the entire image
    - consider the mean and the variance of the output of each filter
  - **Tamura texture features**
    - split images into 9 equal tiles
    - consider contrast, directionality and coarseness
- For HTML:
  - one image node in the representation per *<img>* tag

Cross-Media Analysis (per image)

- Layout distance between sentence and image
  - for HTML: DOM tree analysis

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

- Similarity between sentence and image caption/alt-text

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



- Total cross-media weight:
  - average its components

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

Organisations, Information and Knowledge

Categorization

- Infer category given canonical representation of a document and a previously computed model
- Probabilistic binary one vs all approach

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

- Image-text pairs represented by cross-media feature vector
  - cross-media weight affects text features

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

- Inference with logistic regression:

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

- Model learning:

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

- Class decision for document = average contributions from all images

BBC Web News Dataset

- News articles published by the BBC online
- 6,727 news articles via RSS feed
  - obtained between May and June 2008
- Availability
  - send us an email



Organisations, Information and Knowledge

Dataset Preparation

- Set of corpus-specfic heuristics
- Filter out irrelevant content, e.g.:
  - Advertisements
  - Navigational links

- **Main body identification:**
  - locate DOM tree branch containing the main body of the content
  - handcrafting simple corpus-specific patterns over the XML
- **Noisy-structures removal:**
  - remove remaining videos, comments, navigational links, adverts, etc.
  - again, using corpus-specific patterns
- **Noisy-images removal:**
  - remove images too small to be processed (less than 200 pixels)
  - remove stylistic images (e.g., icons) which have a URL pointing to specific location in the server

Methodology

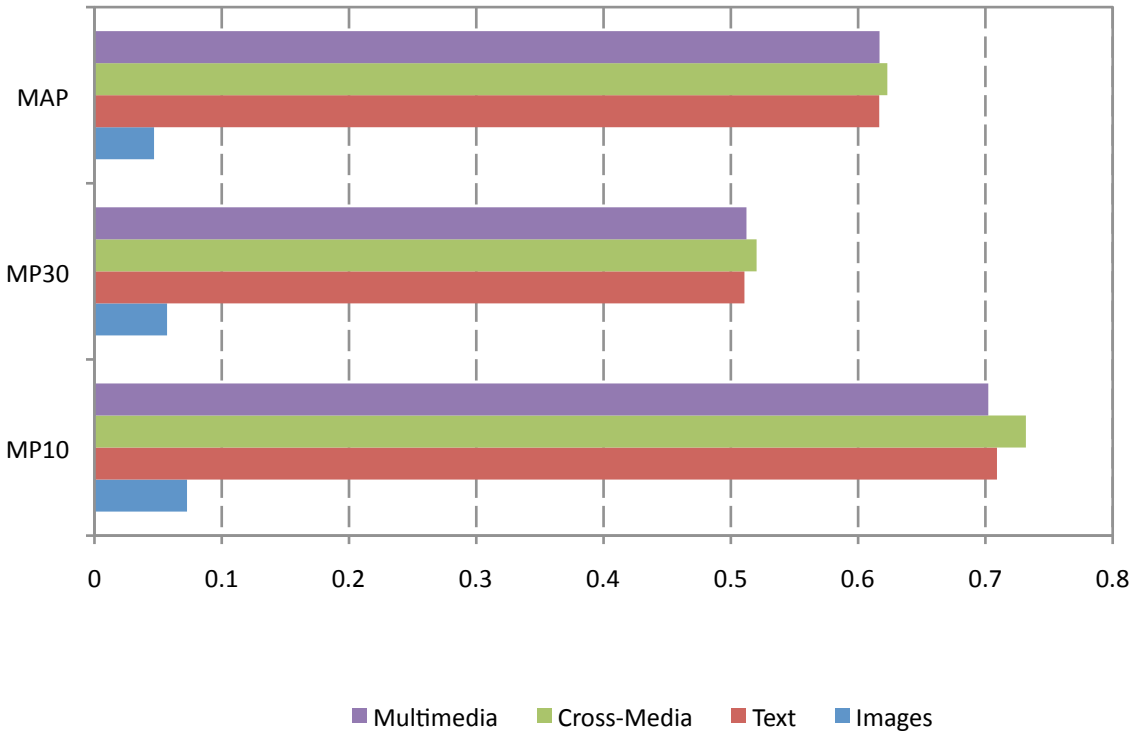
Table 1. BBC news data categories.

/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	

- Single-label classification task
- Labels obtained via the news category assigned by BBC journalists
- Total of 44 categories
- Cross validation:
  - 10 random 2/3 vs 1/3 split
  - (4,485 training vs 2,242 test documents)

Results: Average Precision

- Image results are always much lower
- Cross-media approach yields small but consistently better results

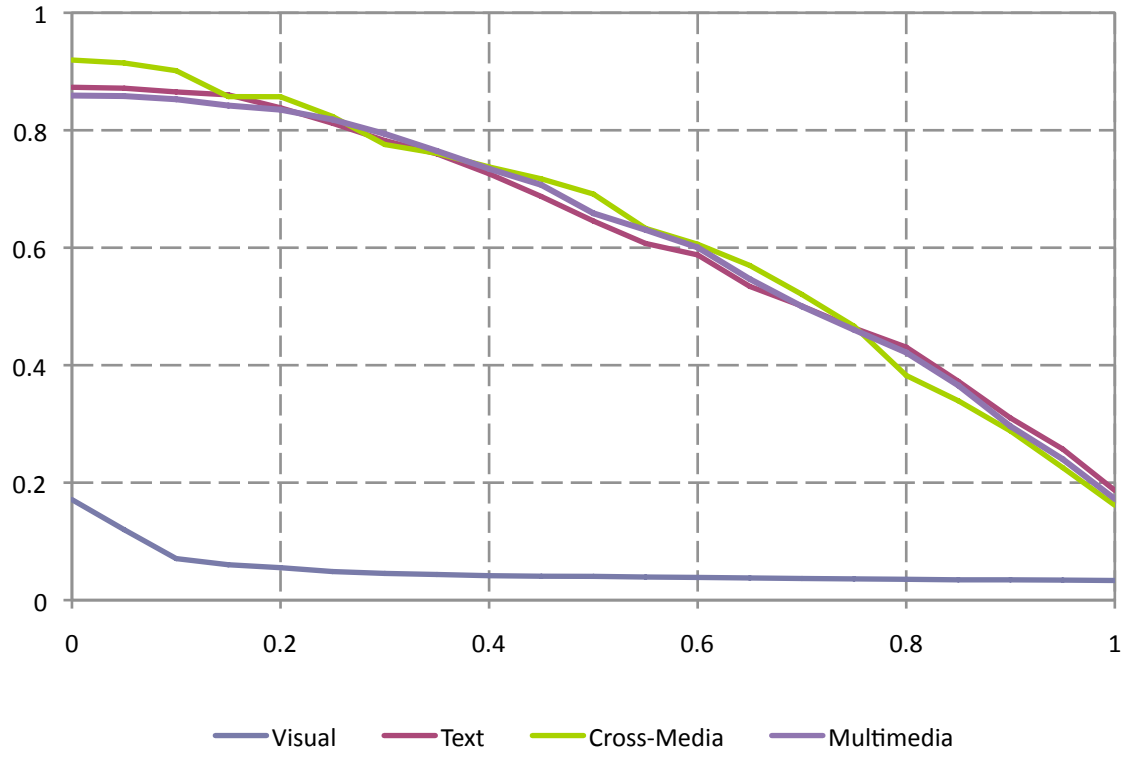


- Simple concatenation of text and image features does not perform much better than text alone

Organisations, Information and Knowledge

Results: Precision-Recall Curves

- Cross-media approach much better at the beginning of the rank (recall < 10%)
  - which is where most users look at



Organisations, Information and Knowledge

Conclusions

- Novel approach to exploiting cross-media correlations from multimedia documents
  - validated in the context of a classification task over multimedia web news documents
- Experimental results empirically show improved system performance
  - by preserving not just text and image features
  - but also measuring the association between text elements and each image
- Advantage of our approach:
  - makes almost no assumptions about the way multimedia content is modeled
  - thus widely applicable