Final POESIA Workshop
"Present and Future of Open-Source Content-Based Web filtering"
Pisa, IT — 21-22 January, 2004

Text Classification for Web Filtering

Fabrizio Sebastiani

Istituto di Scienza e Tecnologie dell'Informazione
Consiglio Nazionale delle Ricerche
56124 Pisa, Italy
fabrizio.sebastiani@isti.cnr.it
http://www.isti.cnr.it/People/F.Sebastiani

Overview of this talk

1. Overview of the Text Classification Task
   (1) A Definition of Text Classification
   (2) Building and Evaluating Text Classifiers
   (3) Applying Text Classifiers
2. Applications of Text Classification to Web Filtering
   (1) Web Classification
   (2) Filtering
   (3) Detecting Unsuitable Content
3. Conclusion

1. Overview of the Text Classification Task

1.1. A Definition of Text Classification

- TC is the task of assigning documents expressed in natural language into one or more categories (aka classes) belonging to a predefined set.
- Mathematically speaking, TC is the task of approximating the unknown target function $\Psi : \mathcal{D} \times \mathcal{C} \rightarrow \{T, F\}$ by means of a function $\Phi : \mathcal{D} \times \mathcal{C} \rightarrow \{T, F\}$ called the classifier, such that Ψ and Φ "coincide as much as possible". Here
  - $\mathcal{C} = \{c_1, \ldots, c_m\}$ is a fixed set of pre-defined categories;
  - $\mathcal{D}$ is a domain of documents.

- We usually make the assumption that the categories are just symbolic labels. No additional knowledge of their meaning is available to help in building the classifier; in particular, the text of which the label consists is not significant.
- The attribution of documents to categories should be realized on the basis of the *content* of the documents (and not on the basis of *metadata* that may be available from an external source).
- Given that the content of a document is a *subjective* notion, this means that the membership of a document in a category cannot be decided with certainty.

- Depending on the application, classification may be
  - Single-label : exactly one category must be assigned to each document. A special case is when $m = 2$ (the binary case).
  - Multi-label : any number of categories can be assigned to each document.
- Depending on the application, a classifier may be required to perform
  - Hard classification, i.e. provide a value in $\{T, F\}$ which indicates membership or non-membership of d_j in c_i. This is useful for autonomous classifiers.
  - Soft classification, i.e. provide a value in $[0, 1]$ which indicates the degree of confidence of the system in the membership of d_j in c_i. This is useful for interactive classifiers.

1.2. Building and Evaluating Text Classifiers

A classifier for $\mathcal{C}$ can be built

- Manually, by knowledge engineering techniques. This means building a set of rules of type

  if ((`wheat` & `farm`) **or**
  (`corn` & `farm`) **or**
  (`wheat` & `commodity`) **or**
  (`wheat` & `tonnes`) **or**
  (`wheat` & `winter` & ¬ `soft`)) **then** GRAIN **else** ¬ GRAIN

- Automatically, by (supervised) machine learning techniques, from a "training" set of documents preclassified under $\mathcal{C}$.

Advantages of the supervised machine learning approach:

- The engineering effort goes towards the construction not of a classifier, but of an automatic builder of classifiers (learner) $\longrightarrow$ if the set of categories is updated, or if the system is ported to a different domain, all that is needed is a different set of manually classified documents.
- Domain expertise (for labelling), and not knowledge engineering expertise, is needed; this is advantageous, since it is easier to characterize a concept extensionally than intensionally.
- Sometimes the preclassified documents are already available.
- The effectiveness achievable nowadays by these classifiers rivals that of hand-crafted classifiers and that of human classifiers.

In the supervised machine learning approach:

- Usually, a document is represented as a (sparse) vector of weights, where the length of the vector is the number of terms that occur in at least one training document.
- Weights may be binary (indicating presence or absence of the term in the document), or non-binary (indicating how much the term contributes to the semantics of the document). In this latter case, weighting functions from text search (such as $tf * idf$) are used.
- Previous to weighting, stop word removal and stemming are often performed. Dimensionality reduction (aka feature selection) may also be performed in order to improve the efficiency of the system.

- Classification is a subjective task, and both human and automatic classifiers are error-prone. The effectiveness of a (human or automatic) classifier is typically measured by comparing its decisions with the human decisions encoded in a "test" set of documents preclassified under $\mathcal{C}$.
- The "classic" effectiveness measure is $F_1 = \frac{2\pi\rho}{\pi + \rho}$, the harmonic mean of *precision* (π) and *recall* (ρ):
  - precision : "How many documents deemed to belong to the category truly belong to it"?
  - recall : "How many documents truly belonging to the category have been deemed as such"?

- Supervised learning techniques often used in ATC are
  - "legacy" techniques: decision trees, decision rules, probabilistic (Bayes) classifiers, neural networks, nearest neighbour techniques, etc.
  - "emerging" techniques: boosting, support vector machines, ridge (regularized) regression.
- In general, any such technique builds a model of the category based on the different distributions that the terms have in the positive and in the negative training examples

  E.g. "`player` appears more often in the positive training examples of Sport that in its negative examples"

1.3. Applying Text Classifiers

Among the applications of ATC we may mention

- document indexing with controlled vocabularies, i.e. classifying documents with categories (or "subject codes", or "descriptors", ...) from e.g. the Library of Congress Cataloging Scheme, the Dewey Decimal System, etc.;
- filing patents under predefined patent categories;
- filing "classified ads" into classes (e.g. deciding whether a given ad should be printed under Cars for Sale, or Real Estate, ...);
- detecting authorship of documents of disputed paternity (e.g. Shakespeare vs. (**not** Shakespeare));
- classifying images through the analysis of textual captions;
- identifying text genre (e.g. ForKids vs. (**not** ForKids)).

2. Text Classification for Web Filtering

- Web Filtering can be seen as combining characteristics of two specific applications of TC, namely
  1. Web Classification , i.e. filing Web sites, or organizing search results, under hierarchical directories (e.g. Yahoo!, DMOZ).
  2. Filtering , i.e. classifying each of a stream of incoming documents into either User_i or (**not** User_i) based on the interest/appropriateness of the document to the user.
- The POESIA application also features characteristics of a third application of TC, namely
  3. Detecting Unsuitable Content (e.g. deciding between Pornography and (**not** Pornography)) or junk mail (Spam vs. (**not** Spam)).

2.1. Web Classification

- Classifying Web Pages is a peculiar case of TC because of the *presence of hyperlinks*. These constitute a rich source of information, as they may be understood as statements of relevance of the linked page to the linking page.
- Research in *bibliometrics* (resp. *hypertext retrieval*, e.g. Google) has already pointed out that the quantitative analysis of bibliographic quotations (resp. hyperlinks) could help in determining relevance of content.

- Research in Web classification has shown that effectiveness can be improved by the use of Web-specific heuristics; e.g. use the categories of the hyper-neighbours as features, and use a "relaxation labelling" iterative technique if these are not known in advance [Chakrabarti+98].
- Web pages are more difficult to analyse than standard text, for several reasons (e.g. use of evocative rather than descriptive language, presence of noise). As a result, it is important to use *sophisticated* techniques in order to reach *reasonable* levels of effectiveness.

2.2. Filtering

- Filtering has two main peculiarities wrt other TC applications
  - False positives and false negatives often have different importance; this must be accounted for by using a utility-theoretic measure of effectiveness.
  - The training set is often acquired incrementally, through the interaction with the user. This means that techniques that generate a classifer incrementally (*on-line learners*) must be used.
- Recently, effectiveness levels comparable to the ones of standard TC applications have been reached in filtering through the use of maximum-margin online learners [Cancedda+02].

2.3. Detecting Unsuitable Content

- Detecting Unsuitable Content is an instance of *Content Management in Adversarial Environments* (CMAE); this refers to the case in which content management applications face the presence of content authors who (usually because of economic interests) do not cooperate with the content seekers. E.g.
  - Unsolicited Bulk Email (aka Spam);
  - "Junk" Web sites.
- The problem with CMAE is that no (filtering or other) technique for CMAE can be considered definitive, since "predators adapt to their prey".

- Filtering pornography and filtering (non-pornographic) spam are two instances of CMAE.
- *Prima facie*, the former is "easier" than the latter, since
  - The distribution of terms within Pornography tends to be very peculiar.
  - Spam cuts across topics, and is thus not easily detected by looking at the distribution of terms. Automatic feature selection is thus usually supplemented with human feature engineering (e.g. check for sentence Make money fast, check if message claims you were on a mailing list, check for ALL CAPS in either subject field or body, etc.).
- However, distinguishing Pornography from SexEd may be less easy. And distinguishing Pornography from Erotica is even harder, since the border is extremely subjective.

3. Conclusion

- Text classification, especially in its machine learning variant, is nowadays an essential technique for Web filtering applications, because of
  - the constraints of the application, that require fast response, adaptivity, no manual work
  - its effectiveness
- For achieving optimum performance in filtering Web documents, Web-specific information should be capitalized upon.
- In filtering unsuitable content, some tasks are more difficult then others, due to subtle and extremely subjective distinctions.
- Utmost care must be taken when acting in "adversarial environments", because of the presence of non-cooperating actors.

And for those interested in text classification ...

- On-line searchable bibliography on ATC (part of the *Collection of Computer Science Bibliographies*), with ≥ 400 entries, at `http://faure.iei.pi.cnr.it/~fabrizio/main.html#ATCbiblio` Most entries are complete with abstracts and URLs to electronic copies.
- Review/tutorial articles:
  - Fabrizio Sebastiani. Machine Learning in Automated Text Categorization. *ACM Computing Surveys* **34**(1):1–47, 2002.
  - Fabrizio Sebastiani. Text Categorization. In Alessandro Zanasi (ed.), *Text Mining and its Applications*, WIT Press, Southampton, UK, 2004. Forthcoming.
  - Fabrizio Sebastiani. Text classification, automatic. In Keith Brown (ed.), *The Encyclopedia of Language and Linguistics*, 2nd Edition, Vol. 14, Elsevier Science, Amsterdam, NL, 2004. Forthcoming.