Experiments on using *Yahoo!* categories to describe documents

Yannis Labrou and Tim Finin
Computer Science and Electrical Engineering Department,
University of Maryland, Baltimore County,
{jklabrou,finin}@cs.umbc.edu
Phone: (410) 455 3624
Fax:(410) 455 3969

Abstract

We suggest that one (or a collection) of names of *Yahoo!* (or any other WWW indexer's) categories can be used to describe the content of a document. Such categories offer a standardized and universal way for referring to or describing the nature of real world objects, activities, documents and so on, and may be used (we suggest) to semantically characterize the content of documents. WWW indices, like *Yahoo!* provide a huge hierarchy of categories (topics) that touch every aspect of human endeavors. Such topics can be used as descriptors the way librarians use for example, the Library of Congress cataloging system to annotate and categorize books.

In the course of investigating this idea, we address the problem of automatic categorization of webpages in the *Yahoo!* directory. We use **Telltale** as our classifier; **Telltale** uses n-grams to compute the similarity between documents. We experiment with various types of descriptions for the *Yahoo!* categories and the webpages to be categorized. Our findings suggest that the best results occur when using the very brief descriptions of the *Yahoo!* categorized entries; these brief descriptions, which are part of the *Yahoo!* index itself accompany most entries. We discuss further research and ways to improve on the performance of our method.

1 Introduction

People are very good at answering the question "what is this about?", where "this" might refer to a book, a newspaper article, a publication, a webpage, *etc.*, especially when "this" falls into an area of human knowledge or experience that they master. Because the beneficiaries of answers to such questions are other people who possess a body of general knowledge and the mastery of a spoken language, they are not troubled by the (often) incomplete and non-standardized nature of the responses. Computer programs on the other hand could benefit from a standardized way for describing the content or the nature of "things" (of all things, we will focus on "things" of a textual form). The descriptions that we have in mind are not semantically deep descriptions of "things" but rather headline-like accounts of their nature that describe them in the broader context of human knowledge and experience. For example, *Phantom of the Opera* might be a `Musical`, or it might be a `Musical`, which is a form of `Theater`, which is a kind of a `Performing Art`, which in turn is something that has to do with the `Arts`; in other words, *Phantom of the Opera* is a `Arts:-Performing Arts:Theater:Musicals` kind of thing.

Librarians have been arduously performing this task for centuries but the emergence of the World Wide Web (WWW) in recent years has led to the creation of huge indices that focus on categorizing selected webpages depending on their content. *Yahoo!* for example, is an attempt to organize webpages into a hierarchical index of more than 150,000 categories (topics). We suggest that a *Yahoo!* category (or a collection of them) can be used to describe the content of a document, the way `Arts:Performing Arts:Theater:-Musicals`, which is indeed a *Yahoo!* subcategory, can be used to refer to *Phantom of the Opera* or to describe a webpage about the musical *Phantom of the Opera*. If eXtended Markup Language (XML) lives up to the high expectations associated with it, one can imagine a tag like `YahooCategory` that can be introduced and supplement the XML source of a webpage, which in effect would describe how this particular webpage could have been categorized in the *Yahoo!* hierarchy.

Such a semantic annotation of documents would be useful, even if it has do be done manually, because it will offer a uniform and universal way of referring to the content of a document. Of course, we need not limit ourselves to document descriptions. Although, for example, agents (human or software ones) can describe their interests, or their capabilities as collections of *Yahoo!* categories, our larger point is that *Yahoo!* categories can be used as a standardized way for referring to or describing the "nature" of things. On the other hand, successfully automating this process offers a whole new array of possibilities. To name a few, it will be easier to classify things into the *Yahoo!* (or any other WWW indexer's) hierarchy, search engines will have an easier task finding things if they are semantically annotated this way, spiders will be able to index a much larger part of the WWW, browsers can be more tuned to their users' particular interests (by tracking accessed documents), and so on.

This paper presents some experiments that explore the automation of the process of semantically annotating webpages via the use of *Yahoo!* categories as descriptors of their content. So, the question we are addressing is: given some random webpage, if a classifier were to categorize it in the *Yahoo!* directory of topics, could it put it at the same place in the hierarchical index that the human indexers of *Yahoo!* would? We are less concerned with the choice of classifier and more interested in identifying the optimal descriptions for the categories and for the webpages to be categorized. Although we use an n-gram based classifier called **Telltale**, we believe that another classifier could have been used for our experiments with possibly better results; our choice was based on the immediate availability of the software and the expertise of its developers. We first discuss some observations about *Yahoo!* (Section 2) that led to our idea to set up these experiments. In Section 3 we present the steps of our experiments. We continue to present our results (Section 4) and to discuss them (Section 5). Before concluding we present our ideas for further research in Section 6.

2 Some observations about *Yahoo!*

Yahoo! is an index of categories (topics), organized in a hierarchical manner. Let us look at the *Yahoo!* page of a particular category. The following is a textual representation of what can be found (or, at least could be found at the time we collected our data) under `http://www.yahoo.com/Arts/Performing_Arts-/Theater/Musicals/`. Category names followed by an "@" are links to other *Yahoo!* categories, classified under a different path of the *Yahoo!* hierarchy (they are like links in the UNIX file-system); so, the *Yahoo!* hierarchy is more like a DAG (directed acyclic graph) than a tree.

```
  ________________________________________________________________

  Top:Arts:Performing Arts:Theater:Musicals
_____________________________ ______Options
___ Search all of Yahoo ___ Search only in Musicals

  * Indices (3)
  ________________________________________________________________

  * Movies@
  * Shows (124) [new]
```

```
* Songwriters@
* Theater Groups (22)
________________________________________________________________
* Australian Musical Theater
* Gilbert and Sullivan@
* Jeff's Musical Page - for Les Miserables, Martin Guerre, and other
  popular musicals.
* Just a Few Miles North of NYC - pictures and clips from favorite
  Broadway shows, original scripts, and a chat room to discuss
  theater.
* MIT Musical Theater Guild Archives - synopses of musicals
* Musical Cast Album Database - searchable database of musicals
  released on compact disc.
* Musical Page - pictures and information from popular musicals. The
  Phantom of the Opera, Sunset Boulevard, and several more.
* Musicals Home Page - an index to many Broadway musicals.
* Rutger's Theatre Gopher
* Tower Lyrics Archive - lyrics for several musicals
* Ultimate Broadway Midi Page - midis from a plethora of Broadway
  shows, as well as librettos and synopsises.
* Wisconsin Singers
* Usenet - rec.arts.theatre.musicals
________________________________________________________________
```

We observe the following items of interest that are present on every *Yahoo!* page describing a *Yahoo!* category (topic) and the chosen entries categorized under this category:

1. First there is a category name which in the above example is:
   `Top:Arts:Performing Arts:Theater:Musicals`

2. Another group of items is the sub-categories of the current category.

   ```
   * Movies@
   * Shows (124) [new]
   * Songwriter@
   * Theater Groups (22)
   ```

   These sub-categories (the children nodes of the current category) come in two varieties: (a) those that point to other categories of the *Yahoo!* hierarchy and are depicted with "@" following their name, and (b) those that are indexed under the current category. So, for the above set of sub-categories, only `Shows` and `Theater Groups` are direct children of `Top:Arts:Performing Arts:Theater:Musicals` and they are going to appear as such in the html document:

   ```
   <a href="/text/Arts/Performing_Arts/Theater/Musicals/Shows/">
   <a href="/text/Arts/Performing_Arts/Theater/Musicals/Theater_Groups/">
   ```

   The other two categories (`Movies@` and `Songwriters@`) as their corresponding URL's suggest, point to other places in the hierarchy

   ```
   <a href="/text/Entertainment/Movies_and_Films/Titles/Musicals/Shows">
   <a href="/text/Entertainment/Music/Composition/Songwriting/Songwriters/">
   ```

3. The most important information is what we can call "semantic content" of the category, in other words the "content" that offers an indirect "description" of the category:

   ```
   * Australian Musical Theatre
   ... other omitted entries ....
   * Ultimate Broadway Midi Page - midis from a plethora of Broadway
     shows, as well as librettos and synopsises.
   ... other omitted entries ....
   ```

Term	Definition
CATEGORY	a particular *Yahoo!* category (topic)
ENTRY	a categorized entry (some non-*Yahoo!* webpage) indexed in a CATEGORY
CATEGORYNAME	the full hierarchical name of a CATEGORY in *Yahoo!*, *e.g.*, `Top:Arts:-Performing Arts:Theater:Musicals`
CATEGORYDESCRIPTION	whatever constitutes the description of the category (see below for elements that can be used in the CATEGORYDESCRIPTION of a CATEGORY)
ENTRYCONTENT	the html document that the ENTRY URL points to; a collection of ENTRYCONTENT descriptions can be used in a CATEGORYDESCRIPTION
ENTRYTITLE	the title of an ENTRY that is often descriptive of the content of the ENTRY, *e.g.*, `Musicals Home Page`; a collection of ENTRYTITLE descriptions can be used in a CATEGORYDESCRIPTION
ENTRYSUMMARY	the brief textual description of an ENTRY that in the case of *Yahoo!* is generated by either the *Yahoo!* classifiers or by the human who submitted the page to *Yahoo!* for indexing, *e.g.*, `an index to many Broadway musicals`; a collection of ENTRYSUMMARY descriptions can be used in a CATEGORYDESCRIPTION

Table 1: Summary of terms and definitions used in this document.

> Every item here is a link outside *Yahoo!* Each entry is presented with a **title**, *e.g.*, `Ultimate Broadway Midi Page`, which could very well be the `title` field from the html document of the page, and is (optionally) accompanied by a **brief description**, *e.g.*, `midis from a plethora of Broadway shows, as well as librettos and synopsises`, which is provided either by the human indexers or by the creator of the webpage when (s)he submitted it to *Yahoo!* for indexing. The latter element of the categorized entries is what we intend to take advantage of.

In Table 1 we summarize various general terms and definitions used in this document. We consider the ENTRIES already categorized under a particular *Yahoo!* category to be the material for the "description" of the category. Our main thesis, is that these ENTRIES provide us with the semantic content of a CATEGORY, in the sense that if a new ENTRY were to be classified under that particular CATEGORY, its content would probably be similar to the content of the ENTRIES already classified under that particular CATEGORY. Our experiments investigate the best way for describing CATEGORIES and ENTRIES. CATEGORIES will be described using combinations of features (ENTRYCONTENT, ENTRYTITLE, ENTRYSUMMARY) of ENTRIES that have **already** been classified. ENTRIES will be described using one of their features (ENTRYCONTENT, ENTRYTITLE, ENTRYSUMMARY). Our goal is to seek the most promising combination of descriptions for CATEGORY and ENTRY.

3 An outline of our experiments

Let us describe the phases of our experiments:

Phase I We replicated the entire *Yahoo!* tree locally (approximately 500 MBytes). Some information relating to the number of *Yahoo!* CATEGORIES and their respective sizes as of the time of the collection of our data can be found in Table 2[1]. By creating a local copy of *Yahoo!*, we could store on our systems all the

[1] An interesting observation is the large number of CATEGORIES that appear to be indexed under the `Regional` top-level CATEGORY (almost 3/4 of all the CATEGORIES).

Top-level CATEGORY	*Number of topics (sub-*CATEGORIES*)*	*Size (in KB)*
Arts	2553	9417
Business and Economy	13401	91551
Computers and Internet	2357	8549
Education	322	1521
Government	3996	27065
Health	1177	4328
News and Media	1617	6728
Recreation	5200	18032
Reference	126	556
Regional	113952	324180
Science	2527	899
Social Science	505	1829
Society and Culture	2797	11255
TOTAL	151763	518510

Table 2: Summary of top-level *Yahoo!* CATEGORIES and their respective sizes.

information necessary for our experiments, without the need for accessing the WWW every time we needed data. We used **Wget**[2], a GNU network utility for retrieving files from the WWW, to download and replicate locally the entire *Yahoo!* hierarchy.

Phase II We generated the CATEGORYDESCRIPTION and the test cases (from here-on referred to as TESTCASES). In Section 2 we mentioned that there are a number of elements that we can choose to construct the CATEGORYDESCRIPTION. For the round of experiments described here, we decided on three TYPES of CATEGORYDESCRIPTION: TYPEA, TYPE and TYPEC (see Table 3). We also had to make similar decisions regarding the test cases to be used in the experiments (the test cases were ENTRIES that were already categorized in *Yahoo!*). We used three different ways to describe them: TESTCASETYPEA, TESTCASETYPEB and TESTCASETYPEC (see Table 3).

The chosen TESTCASES were removed, *i.e.*, were not accounted for as ENTRIES when constructing the various CATEGORYDESCRIPTIONS. We used some simple heuristics in order to ensure an even distribution of a sufficient number of TESTCASES across the entire collection of *Yahoo!* CATEGORIES (basically, we took into account the density of each top-level CATEGORY and we tried to allocate same order of magnitude numbers of TESTCASES for each top-level CATEGORY).

Phase III We generated the corpus and ran the experiments.

We used **Telltale** as our classifier. **Telltale** [11, 3, 2] was developed at the LABORATORY for ADVANCED INFORMATION TECHNOLOGY, at the CSEE Department of UMBC; among other things, **Telltale** can compute the similarity between documents, using n-grams as index terms. The weight of each term is the difference between the count of a given n-gram for a document, normalized by its size, and the average normalized count over all documents for that n-gram. This provides a weight for each n-gram in a document relative to the average for the collection (corpus). The similarity between documents is then calculated as the cosine of the two representation vectors.

Our goal was to generate a single corpus of all *Yahoo!*categories and the to run our experiment for

[2] `http://www.lns.cornell.edu/public/COMP/info/wget/wget_toc.html`

CATEGORYDESCRIPTION types	TYPEA	the collection of the ENTRYSUMMARIES and ENTRYTITLES for each ENTRY of a given CATEGORY
	TYPEB	the collection of the ENTRYSUMMARIES for each ENTRY of a given CATEGORY
	TYPEC	the combination of TYPEA and the CATEGORYNAME, *i.e.*, the collection of the ENTRYSUMMARIES and ENTRYTITLES for each ENTRY of a given CATEGORY along with the CATEGORYNAME of the CATEGORY.
TESTCASES types	TESTCASETYPEA	we use the ENTRYTITLE of the ENTRY
	TESTCASETYPEB	we use the ENTRYSUMMARIES of the ENTRIES
	TESTCASETYPEC	we use the ENTRYCONTENT of the ENTRY; we were careful to only select as TESTCASES those ENTRIES that pointed to URLs that contained a sufficient amount of text (file size bigger that 1K discounting images, imagemaps, soundfiles, *etc.*)

Table 3: Summary of terms and definitions related to the experiments.

	CATEGORY	TESTCASES
EXPERIMENTA	`Education` and `Social Sciences`	from `Education`
EXPERIMENTB	`Education` and `Social Sciences`	from `Social Sciences`
EXPERIMENTC	`Health`	from `Health`

Table 4: The three experiments we conducted

each one of TYPEA, TYPEB and TYPEC and for every set of TESTCASES of each type (TESTCASETYPEA, TESTCASETYPEB and TESTCASETYPEC), for a total of 9 experiments (one for each combination of CATEGORYDESCRIPTION and ENTRYDESCRIPTION). For each experiment we expected to compute the similarity of each TESTCASE type and CATEGORYDESCRIPTION type, order them in descending order (using some cut-off point for the similarity measure) and finally returne the **position** of the **correct match**; the **correct match** is the CATEGORY under which the TESTCASE was actually classified in *Yahoo!* before being removed for the experiments.

4 Experimental Results

When we started **Phase III** we realized that **Telltale** was not up to the task of generating the huge corpuses we needed for the experiments. Merging the corpuses of each of the top-level CATEGORIES into a single *Yahoo!* corpus proved to be an insurmountable obstacle. Since the new version of **Telltale** was under way we decided to modify our immediate goals and to postpone the full version of our experiment until the new and improved implementation of **Telltale** became available. Instead of checking the test cases against the entire collection of CATEGORIES (a single corpus) we decided to run 3 experiment sets, with different combinations of top-level CATEGORIES (so we generated 3 corpora of *Yahoo!* categories instead of one) and TESTCASES. More specifically, in each of these experiment sets, the TESTCASES were drawn from a different top-level *Yahoo!* category and matched against CATEGORIES from a single top-level CATEGORY (*i.e.*, `Health`) or a combination of such (*i.e.*, `Education` and `Social Sciences`), as summarized in Table 4.

	EXPERIMENTA					
	TESTCASETYPEA		TESTCASETYPEB		TESTCASETYPEC	
	1	1-10	1	1-10	1	1-10
TYPEA	22 (**37%**)	38 (**64%**)	35 (**63%**)	45 (**82%**)	13 (**50%**)	18 (**69%**)
TYPEB	16 (**27%**)	30 (**51%**)	16 (**29%**)	34 (**62%**)	9 (**35%**)	14 (**54%**)
TYPEC	20 (**34%**)	38 (**64%**)	23 (**42%**)	34 (**62%**)	8 (**29%**)	15 (**54%**)

Table 5: Results from EXPERIMENTA; the corpus is comprised from the top-level CATEGORIES of `Education` and `Social Sciences` and the TESTCASES are drawn from `Education`. We provide the absolute numbers and the percentages of the TESTCASES that were returned in the top position (1) of the list returned and in the (1-10) range.

	EXPERIMENTB					
	TESTCASETYPEA		TESTCASETYPEB		TESTCASETYPEC	
	1	1-10	1	1-10	1	1-10
TYPEA	11 (**20%**)	25 (**46%**)	37 (**82%**)	44 (**98%**)	8 (**40%**)	17 (**85%**)
TYPEB	7 (**13%**)	20 (**37%**)	10 (**22%**)	25 (**56%**)	4 (**20%**)	12 (**60%**)
TYPEC	16 (**30%**)	24 (**44%**)	23 (**51%**)	36 (**80%**)	7 (**35%**)	13 (**65%**)

Table 6: Results from EXPERIMENTB; the corpus is comprised from the top-level CATEGORIES of `Education` and `Social Sciences` and the TESTCASES are drawn from `SocialSciences`. We provide the absolute numbers and the percentages of the TESTCASES that were returned in the top position (1) of the list returned and in the (1-10) range.

For each one of EXPERIMENTA, EXPERIMENTB and EXPERIMENTC, we ran 9 experiments, one for each combination of CATEGORYDESCRIPTIONS (TYPEA, TYPEB and TYPEC) and TESTCASES types (TESTCASETYPEA, TESTCASETYPEB and TESTCASETYPEC), for a total of 27 experiments. For each of the 27 experiments we return 2 results: **(1)** the percentage (and absolute number of test cases) of times that the **correct match** appeared **first** in the list returned by **Telltale**, and **(2)** the percentage (and absolute number of test cases) of times that the **correct match**, appeared in one of the **first ten** positions in the list returned by **Telltale**. Table 5 shows the results for all 9 experiments for EXPERIMENTA; likewise for EXPERIMENTB and EXPERIMENTC in Table 6 and Table 7, respectively. Finally, in Table 8 we present the averages across experiments EXPERIMENTA, EXPERIMENTB and EXPERIMENTC.

After evaluating the results we can draw the following conclusions: **(1)** The most successful combination of CATEGORYDESCRIPTION and ENTRY descriptions is TYPEA with TESTCASETYPEB, *i.e.*, choosing the collection of the ENTRYSUMMARIES and ENTRYTITLES for each ENTRY of a given CATEGORY to describe the CATEGORY and choosing the ENTRYSUMMARY of the ENTRY to describe the ENTRY **(2)** TYPEA outperforms all the other CATEGORYDESCRIPTIONS, regardless of the choice of entry description, and **(3)** TESTCASETYPEB outperforms all the other entry descriptions, regardless of the choice of CATEGORYDESCRIPTION.

5 Discussion

The purpose of this experiment was to automatically categorize web-documents in the *Yahoo!* hierarchy. Researchers in the areas of Machine Learning and Information Retrieval have experimented with catego-

	EXPERIMENTC					
	TESTCASETYPEA		TESTCASETYPEB		TESTCASETYPEC	
	1	1-10	1	1-10	1	1-10
TYPEA	46 (**37%**)	75 (**60%**)	90 (**75%**)	114 (**95%**)	30 (**43%**)	55 (**80%**)
TYPEB	32 (**26%**)	71 (**57%**)	30 (**30%**)	78 (**65%**)	21 (**30%**)	41 (**59%**)
TYPEC	56 (**45%**)	81 (**65%**)	60 (**50%**)	88 (**73%**)	19 (**29%**)	46 (**70%**)

Table 7: Results from EXPERIMENTC; the corpus is comprised from the top-level CATEGORY of `Health` and the TESTCASES are drawn from `Health`. We provide the absolute numbers and the percentages of the TESTCASES that were returned in the top position (1) of the list returned and in the (1-10) range.

	All Experiments					
	TESTCASETYPEA		TESTCASETYPEB		TESTCASETYPEC	
	1	1-10	1	1-10	1	1-10
TYPEA	31%	57%	73%	92%	44%	78%
TYPEB	22%	48%	27%	61%	28%	58%
TYPEC	36%	58%	48%	72%	31%	63%

Table 8: Averages of the percentages from the results from EXPERIMENTA, EXPERIMENTB and EXPERIMENTC.

rization into hierarchical indices. But our experiments are not comparable with the ones described in [6] and [14], for example, because of the difference in the order of magnitude of the number of categories (less than 20 in [6], more than 1000 in our case) that we are attempting to match against. A fair evaluation of the results has to take into account the sheer number of categories been considered when a webpage is evaluated for classification.

The only similar work we are aware of [3] is the *Yahoo Planet* project [8, 10] which uses the *Yahoo!* hierarchy of Web documents as a base for automatic document categorization. Several top categories are taken as separate problems, and for each an automatic document classifier is generated. A demo version of the system[4] enables automatic categorization of typed text inside the sub-hierarchy of a selected top *Yahoo!* category. Users can categorize whole documents by simply copying their content into a window and requesting categorization of the "typed" text. Their methodology differs in that they built a classifier for each category which learns from positive (correctly indexed webpages) and negatives examples; unlike our method, they do not make use of the brief summaries of the categorized entries. This work relies on Machine Learning techniques and is part of a much larger endeavor [9]. In terms of comparing the results, one should keep in mind two basic differences: **(a)** a top level *Yahoo!* category has to be pre-selected (in experiments EXPERIMENTA and EXPERIMENTB we use a combination of two top-level categories), and **(b)** their metric is slightly different than ours, *i.e.*, they present the median of the correct category, *e.g.*, a result of "median of rank of correct category" equal to 3, means that half of the testing examples are assigned a rank of 1, 2 or 3 [5]. In their experiments the medians for the top-level categories of `References`, `Education` and `Computers and Internet`, are 2, 3 and 3 respectively. By comparison, the results of Table 5, where the test cases are drawn from `Education` and matched against the *combined* top-level categories of `Education` and `Social Sciences` suggest a median of 1 (since 50% of the test cases have a rank of 1), for the case of TESTCASETYPEC (which is equivalent to their "description" of the test case). But

[3]We were not aware of this work, at the time we conceived and ran our experiments.

[4]`http://ml.ijs.si/yquint/yquint.exe`; it does not seem to be running anymore.

again, an one-on-one comparison is impossible. We only consider test cases that have enough text in them and although they also employ similar criteria to make sure that a webpage has enough text to work with, any comparison will be incomplete and inaccurate unless we attempt to categorize exactly the same set of test cases.

If a webpage (or documents) were to be classified automatically one would expect 100% accuracy by the classifier. In that sense, ours is a failed experiment. With respect to a fully automatic categorization of webpages, our approach presents an additional shortcoming: the best performance occurs when some brief textual description of the webpage is used, as is the case with most of the webpages categorized in *Yahoo!*. If webpages are to be categorized without human intervention, no such brief description is expected to be provided. It is quite surprising though, how encouraging the results are even when just a few words are available. On the other hand, it seems that the collection of ENTRYSUMMARY and ENTRYTITLES (TYPEA) is extremely powerful in terms of describing the content of a particular CATEGORY. An observation in favor of our results is that we take the *Yahoo!* indexing to be the absolute and only correct categorization of a document. In other words, we do not investigate whether the matches returned by our classifier are reasonable or correct matches, even if the *Yahoo!* indexers thought otherwise (perhaps because of the additional time needed to classify a webpage in multiple locations in the hierarchy). Finally, we discount as false a result that returns a CATEGORY that even though is not the correct one is pretty close (semantically) to it.

Maybe a proper evaluation of the results depends on the potential use of this technique. Inadequate as it might be for a strictly automated categorization of webpages, it could be useful for offering suggestions to a human indexer. If though, the owner of the webpage is willing to provide a very brief account of the webpage, our method could be useful for automatic categorization. Finally, if the method is used for automatically tagging webpages (or documents) in order to semantically describe their content, the error might be within acceptable range for the purpose.

6 Future work

Our next goal is to experiment with the new version of **Telltale** which will allow us to test the TESTCASES against a corpus of all the *Yahoo!* topics minus the `Regional` category (a total of approximately 38,000 categories). One of our observations about *Yahoo!* is that 3/4 of its topics are indexed under the `Regional` top-level category. It seems that most of the topics indexed somewhere in a `Regional` sub-category could have also been indexed under another top-level category but they do not appear there too. For example, imagine some small-town real estate agency which is indexed under the real-estate businesses of the small town's CATEGORY under `Regional` but not under real-estate businesses, under the top-level `Business and Economy` category. Our experiments so far dealt with only 1000 topics, so we do not know what to expect after a one to two orders of magnitude increase. Intuitively, we expect that our current results constitute a best-case upper bound for future results.

Another direction for future experimentation would be to experiment with other classifiers. We used n-grams and **Telltale** because the system was readily available to us and we had immediately available expertise on how to use it for our purposes. We want to experiment with a term frequency/inverse document frequency (TF/IDF) weighting scheme for **Telltale**; [7] suggests that TF/IDF outperforms the centroid weighting method that **Telltale** currently employs. It would also be worth investigating classifiers that take into consideration the hierarchical structure of the *Yahoo!* topics, a future that we did not explore in our experiments. We would like to improve the performance of the TESTCASETYPEC type of an entry's description. This would be crucial if we were to use the technique for automatic categorization, since in this case we can only rely on the html content of the web-document (or the text of a document, in general). So far, our approach with the html content was very basic. Other than making sure that there was enough textual content present, we did not further manipulate its content.

Finally, we would like to re-consider the evaluation of the matches returned by the classifier. Some of the top matches might not be the "perfect" match, *i.e.*, the official *Yahoo!* categorization of a given webpage but they might be close enough to the perfect match in the huge *Yahoo!* DAG, to be useful for providing some sort of semantic information about the content of the webpage (less accurate but still useful). Also, besides considering such "approximate" matches, it would be interesting to have people evaluate the results returned by the classifier. Just because a webpage was classified by the *Yahoo!* human indexers in a particular category, this does not mean that other possible **correct** categories do not exist, some of which might have been returned by our classifiers. So, we would like to have human indexers evaluate the accuracy of the returned matches without knowledge of which match might have been the *Yahoo!* one. We want to re-evaluate the performance of our method under such revised metrics.

7 In conclusion

In this paper we presented a claim and a set of experiments. The claim was that one could use the pre-defined categories of one of many WWW indexers to describe the nature or the content of "things". Although of all "things" we focused on documents we believe that such categories can be used to describe a large range of activities, objects, *etc.* Our experiments and the success thereof is independent of the claim, which by itself we did not validate feeling that the usefulness of such a standardized way of referring to or describing "things" is rather obvious for computer applications. Our experiments investigated the automation of the process of finding the correct description, *i.e.*, a WWW indexer's category (specifically a *Yahoo!* category), to describe a particular kind of "thing", *i.e.*, a webpage. In principle, little would change if instead of a webpage we had chosen a document that focuses on some particular topic. Our results indicated that the specific method we used (using a classifier called **Telltale**) cannot be used alone to automatically categorize documents, if the actual text of the document is used for the classification. One of our main observations though was that a very brief description of the document dramatically improves the effectiveness of the classification. So, given our working assumption that automatic classification would require almost a 100% accuracy we believe that the best use of our method would be in conjunction with a human to which our classifier would offer recommendations. One other important result was that the collection of the brief summaries that accompany the indexed (under a particular category) webpages in *Yahoo!* are extremely useful in capturing what a category is about. This result might be of interest to other researchers interested in similar problems.

References

[1] Grace Crowder and Charles Nicholas. An approach to large scale distributed information systems using statistical properties of text to guide agent search. In Tim Finin and James Mayfield, editors, *CIKM'95 Workshop on Intelligent Information Agents*, 1995.

[2] Grace Crowder and Charles Nicholas. Resource selection in cafe: an architecture for networked information retrieval. In *SIGIR'96 Workshop on Networked Information Retrieval*, ETH Zurich, 1996.

[3] Grace Crowder and Charles Nicholas. Using statistical properties of text to create metadata". In *First IEEE Metadata Conference*, 1996.

[4] Grace Crowder and Charles Nicholas. Meta-data for distributed text retrieval. In *SIGIR'97 Workshop on Networked Information Retrieval*, Philadelphia, 1997.

[5] M. Grobelnik and D. Mladenic. Efficient text categorization. In *ECML-98 Workshop on Text Mining*, 1998.

[6] Daphne Koller and Mehran Sahami. Hierarchically classifying documents using very few words. In *Proceedings of the Fourteenth International Conference on Machine Learning (ICML-97)*, pages 170–178, 1997.

[7] James Mayfield and Paul McNamee. N-grams vs. words as indexing terms. In *TREC-6 Conference Notebook Papers*, 1997.

[8] Dunja Mladenic. Feature subset selection in text-learning. In *Proceedings of the 10th European Conference on Machine Learning ECML98*, 1998.

[9] Dunja Mladenic. *Machine Learning on non-homogeneous, distributed text data*. PhD thesis, University of Ljubljana, Slovenia, October 1998.

[10] Dunja Mladenic. Turning yahoo into an automatic web-page classifier. In *Proceedings of the 13th European Conference on Artificial Intelligence ECAI98*, pages 473–474, 1998.

[11] Claudia Pearce and Ethan Miller. The telltale dynamic hypertext environment: Approaches to scalability. In James Mayfield and Charles Nicholas, editors, *Advances in Intelligent Hypertext*, Lecture Notes in Computer Science. Springer-Verlag, 1997.

[12] Claudia Pearce and Charles Nicholas. Using n–gram analysis in dynamic hypertext environments. In *Proceedings of the Second International Conference on Information and Knowledge Management (CIKM '93)*, 1993. (This paper was also released as UMBC technical report CS –93–10.).

[13] Claudia Pearce and Charles Nicholas. Telltale: Experiments in a dynamic hypertext environment for dynamic and degraded data. *Journal of the American Society for Information Science*, April 1996.

[14] Erik Wiener, Jan O. Pedersen, and Andreas S. Weigend. A neural network approach to topic spotting. In *Symposium on Document Analysis and Information Retrieval*, pages 317–332, Las Vegas, NV, 1995. ISRI; Univ. of Nevada, Las Vegas.