Surface features in video retrieval

Thijs Westerveld, Arjen P. de Vries and Georgina Ramírez

CWI, INS1, PO Box 94079, 1090 GB Amsterdam, The Netherlands

Abstract. This paper assesses the usefulness of surface features in a multimedia retrieval setting. Surface features describe the metadata or structure of a document rather than the content. We note that the distribution of these features varies across topics. The paper shows how these distributions can be obtained through relevance feedback and how this allows for adaptation of (content-based) search results for topic or user preference. An analysis of the distribution of surface features in the TRECVID collection indicates that they are potentially useful, and a preliminary feedback experiment confirms that exploiting surface features can improve retrieval effectiveness.

1 Introduction

Multimedia retrieval typically relies on either low-level content features, like colour or texture descriptors, or on collateral text, like manual annotations or speech transcripts. A third source of information is however often overlooked, namely the *surface features*. Surface features are those properties of (multimedia) documents that do not describe their content. Examples include, the length of a document, a reference to where the document is located, and the production date of a document. Although these features do not directly relate to the document's content, they can be valuable additional sources of information in a retrieval setting. In text retrieval for example, the length of a document is often used as an indicator of relevance (longer documents are more likely to be relevant). Similarly, the number of hyperlinks pointing to a document is an indicator of the importance of a document [1, 2]. Also, in the design of video browsing interfaces, the importance of surface features, like the temporal structure of video, is well-known, see for example [3]. In video search systems however, surface features are mostly ignored.

This paper assesses the usefulness of a number of surface features for multimedia retrieval. We take the TRECVID2004 collection, a collection of CNN and ABC news broadcasts from 1998, as a case study. The rest of this paper is organised as follows. The next section discusses briefly the collection and its surface features. Section 3 studies the distribution of surface features in relevant documents. Section 4 discusses how to acquire information about these distributions in practise. Section 5 shows how the knowledge about surface feature distributions can be used in a retrieval setting. We conclude the paper with a discussion of experimental results.

2 Surface features in the TRECVID collection

TRECVID is a workshop series with the goal of promoting progress in content-based retrieval from digital video via open, metrics-based evaluation. This paper focuses on TRECVID's search task, defined as follows:

> Given the search test collection, a multimedia statement of an information need (topic), and the common shot boundary reference for the search test collection, return a ranked list of at most 1000 common reference shots from the test collection, which best satisfy the need.

The TRECVID2004 test collection consists of 70 hours of ABC and CNN news broadcasts from 1998. The collection is shot segmented and comes with a predefined set of keyframes. The 25 topics in the test collection are multimedia descriptions of an information need, consisting of a textual description and one or more image or video examples. For each topic, relevance judgements are available; these indicate which shots are relevant for the topic.

From the metadata associated to videos and shots, the following surface features can be extracted: the broadcaster, the date of broadcast, the time of the shot within the video, and the duration of the shot.

The TRECVID workshop prohibits exploiting the knowledge that the broadcasts in the collection are from the second half of 1998. This means for example that we cannot directly infer that if someone is looking for shots of Bill Clinton, it would be helpful to include the term *impeachment* in the query. The rationale is that this would be unrealistic. We think however it *is* realistic for a user to have some knowledge of the collection they are searching. It is indeed questionable whether such knowledge is available at system development time, since that would mean that a separate system has to be built for each new data set. Still, the information could be deduced from co-occurrence patterns in known relevant documents that are obtained through (blind) relevance feedback (Section 4).

3 Analysis of the distribution of surface features

We start with an analysis of the various surface features and their distribution in relevant documents and in the collection as a whole. The statistics reported in this section are based on knowledge of the full set of relevant documents.

3.1 Local clustering of relevant shots

A first observation is that relevant shots tend to cluster: when a shot is relevant for a given topic, it is likely that its neighbouring shots are relevant as well. An explanation for this is that the news broadcasts are organised in stories. A story typically shows multiple shots related to the same subject. Thus, when a topic is directly related to a news event, it is obvious that all, or at least many, shots from the story are relevant (see Figure 1). Of course, in a visual retrieval task, topics do not always ask for news events, but even there the relevant shots

tend to cluster with one or more news event, because news stories are typically sequences of (alternating) shots in a given location or situation. When one of the shots shows a relevant item, it is likely to re-appear in other shots of the same story. For example, a news story that happens to be shot on a rainy day, is probably a good source of information when one is searching for shots with umbrellas (see Figure 2).

Fig. 1. *Floods*; relevant images are directly related to a news event (five consecutive shots are shown)

Fig. 2. *Umbrellas*; relevant images cluster with news events (five consecutive shots are shown)

The distribution of the distance from a relevant shot to the next (Figure 3) shows that for almost half of the relevant shots, at least one of the neighbouring shots is also relevant. The histogram shows the average over all TRECVID2004 topics, for some topics up to 80% of the relevant shots have a relevant neighbour. Clearly, relevant shots tend to cluster.

It is debatable whether these clusters should be treated as separate relevant items. When the information need is related to a news subject, perhaps the shots should be grouped, and a sequence of related items should be presented as a single news story. However, when the information need is visual, and shots are judged on their appearance, each of them can be treated as a separate result.

3.2 Video specific features

Another source of information is the metadata associated with the videos, like the date of broadcast or the broadcaster. Note that these features relate to a whole video rather than to individual shots, thus, they can never distinguish between shots from the same video. Still, they may give information as to where in the collection to search for a given topic.

We studied the date of the broadcasts and found that for some topics relevant shots —or rather, broadcasts containing relevant shots— cluster in time. This

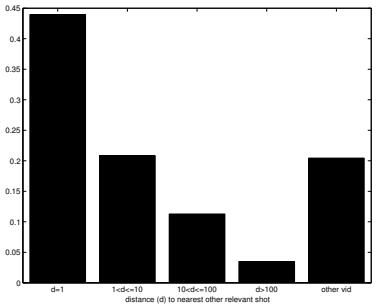

Fig. 3. Histogram of distance to nearest other relevant shot over all relevant shots. Distance measured in number of shots.

is mainly the case when the topics are directly related to news events, like for the examples shown in Figure 4: *floods* mainly occurred in late October, early November 1998; Henry Hyde was the lead house manager in the Clinton impeachment trial that was televised a lot in late December 1998. We also studied the distribution over the different days of the week. For this collection and this set of topics it appears to be random, but one could imagine topics and collections where also the day of the week of a broadcast can be a valuable source of information. For example, there may be more sports news during the weekend, and film related news may cluster on Thursdays or Fridays, when new films are opening.

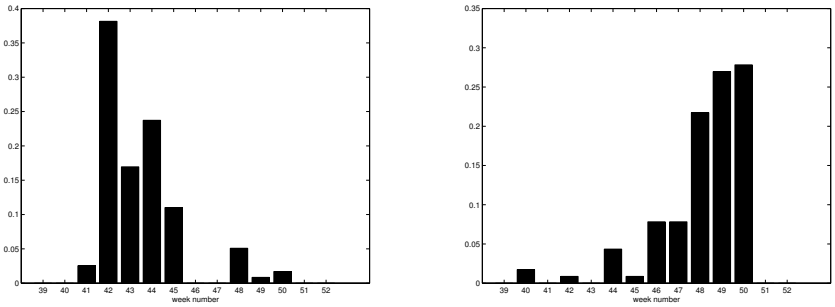

Fig. 4. Distribution of relevant shots over time for two topics: *floods* (left) and *Henry Hyde* (right)

A second attribute associated to the videos is its broadcaster. For many topics the distribution of relevant shots over the two broadcasters in the collection (ABC and CNN) is not uniform. An obvious example is the topic asking for *Sam Donaldson*, a reporter for ABC, and therefore exclusively found in ABC broadcasts. Other topics show perhaps less obvious differences. For example,

relevant shots for sports related topics (*hockey rinks* and *golf*) are mainly found in CNN videos, while the topics *Saddam Hussein* or *buildings on fire* have the majority of relevant shots in ABC videos.

3.3 Time within video

The minute at which a shot starts may also be an indicator of relevance for a given topic. Figure 5 shows the distribution of starting minute over shots in relevant documents compared to its distribution in the collection. On average, relevant information seems to appear more at the beginning of videos (again this is probably because of the news oriented nature of some topics). It may be somewhat surprising that shots in the collection are not uniformly distributed over the minutes. In the first 10 minutes of the videos fewer shots start than in the minutes after that. This indicates that the early shots in the broadcasts are longer. The peaks around 14, 20 and 24 minutes correspond to commercial breaks, typically composed of many very short shots. These commercial break peaks are even more pronounced in the distribution for CNN only (Figure 6). The more uniform distribution in the ABC videos can be explained from the fact that on average the shots in news ABC are slightly shorter (5.7 seconds vs. 6.6 seconds in CNN videos).

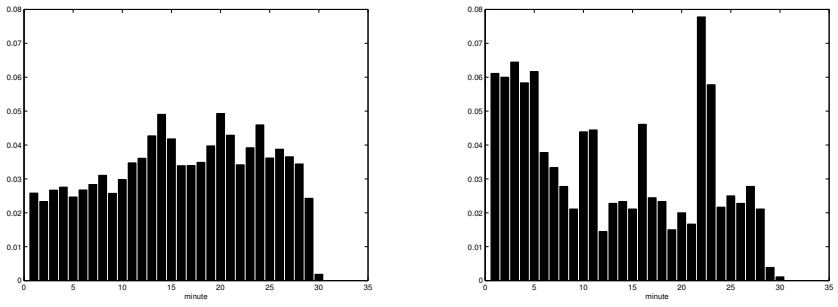

Fig. 5. Distribution of shots over minutes in collection (left) and in set of relevant documents (right).

From Figure 5 we learn that relevant shots in general are likely to appear early in the videos. When we differentiate for individual topics, we may find very different patterns though. For example, the distribution of relevant shots for topics directly related to major news events (*floods*, *Clinton in front of US flag*) have an even higher peak at the beginning of videos (Figure 7). Relevant shots for sports related topics are found between minutes 20 and 25 (Figure 8).

We also analysed the duration of shots, but there seem to be no topics for which the duration of relevant shots clearly differs from the duration of non-relevant shots.

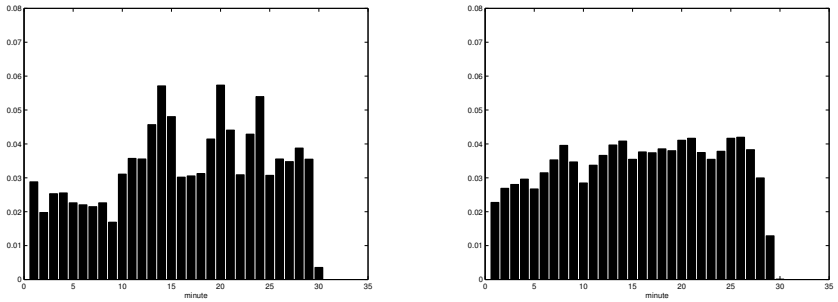

Fig. 6. Distribution of shots over minutes in CNN videos (left) and in ABC videos (right).

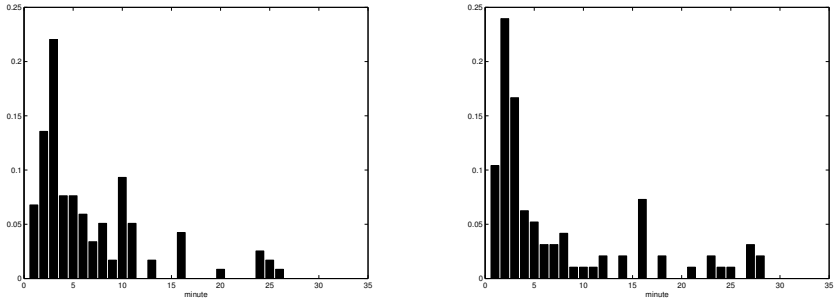

Fig. 7. Distribution over minutes of shots relevant to *floods* (left) and of shots relevant to *Clinton with US flag*

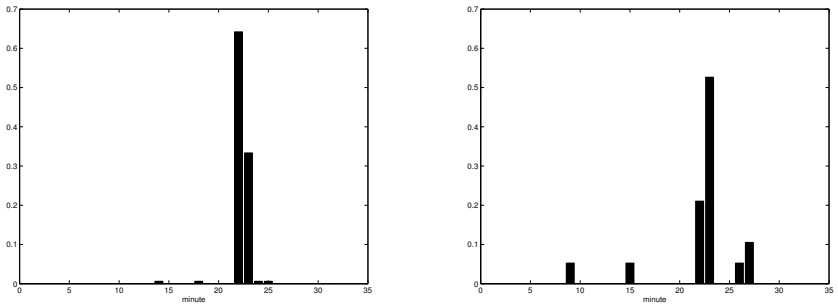

Fig. 8. Distribution over minutes of shots relevant to *hockey rinks* (left) and of shots relevant to *golf score*

4 Acquiring surface feature statistics

In the previous section, we saw how surface features are potentially useful in search. Their distribution distinguishes relevant shots from non-relevant shots (especially when the features are analysed on a topic-by-topic basis). This section discusses how to obtain information about the surface feature distributions for a given topic.

4.1 User-defined priors

Since the user is likely to have some intuition about where relevant items appear, a simple approach would be to let the user explicitly define the surface features that are expected on relevant documents. For example, a user familiar with the data set should be able to tell the system that relevant items for sports related topics mainly appear after some 20 minutes in the news broadcast.

4.2 Topic classification

An alternative would be to develop a taxonomy of topic types and their corresponding surface feature distributions. Such a taxonomy could for example include separate classes for sports, politics and disasters; in addition, it could include knowledge about the broadcasters, like lists of reporters. Topics could be classified manually by the user, who no longer needs to know the relevant feature distributions as was required in the previous subsection. Alternatively, the class could be deduced automatically from the topic description using thesauri like WordNet. The corresponding surface feature distributions could be used to boost the scores of shots with the preferred features.

Some prior work exists on classifying topics and treating each topic differently [4, 5], there the classification is used to decide how to combine the various retrieval modalities (e.g., ASR, visual). An approach to use a surface feature, the time within the video, for searching for topics classified as *weather news* or *sports event* is discussed in [6]. More research is needed to investigate how well the approach extends to other topic types.

4.3 Relevance feedback

The approaches discussed in the previous subsections only work when the topic is clearly related to a news item. One cannot expect users —let alone systems— to be able to guess which major news events coincide with shots of say *umbrellas* or *wheelchairs*. For such topics, the direct modelling of topic classes will not work.

An alternative method uses relevance feedback, and deduces the desired surface feature statistics from shots that are known to be relevant. This information can be obtained from the user who judges the top N documents of an initial retrieval step. Alternatively, we could use blind feedback and assume all of the

top K documents are relevant. The distribution of the surface features in the set of relevant documents can be analysed, and shots with similar distributions can be preferred in the next retrieval round.

Blind feedback may be problematic in a visual retrieval setting, since the effectiveness of present content-based image retrieval systems is limited. Assuming that the top retrieved documents are all relevant is risky. Alternatively, this feedback step could be performed on a comparable text corpus (e.g., news paper articles from the same period). Correspondence of query terms to major news events or clustering of relevant documents in time can be found in this text corpus and transfered to the multimedia collection to improve retrieval there (the same technique has been applied to query expansion using Google news [5]). Admittedly, searching in textual corpora limits the exploitation of surface features to news related topics, because it is unlikely that visual characteristics that happen to coincide with the news event (e.g., it is a rainy day; people carry umbrellas) are mentioned in the papers.

5 Experiments

On their own, surface features may be of little use, but they may improve results obtained from traditional text-based or content-based retrieval methods. The basic idea is to run an ordinary retrieval system based on textual or visual features and then update the scores based on the surface features. Shots with surface features that are likely to be relevant for a given topic will be pushed up the ranked list. This section discusses preliminary experiments with this approach. The experiments reported on here start from a text based retrieval system, but could easily be extended to a content-based setting.

We use a language modelling approach to information retrieval [7, 8]. This retrieval model allows for easy integration of information gathered from surface features. All that is needed is an estimate of the prior probability of relevance given a particular surface feature (or a set of such features). The content scores can then easily be updated, simply by multiplying them with this prior (similar techniques in web retrieval and XML retrieval are discussed in [9, 10] respectively). Note that in other retrieval models, surface features can be incorporated in a similar fashion by using a weighting of returned element scores based on their surface features.

All experiments described in this section are based on text only runs. The shots in the collection are described by the ASR transcripts provided by LIMSI [11]. The score of a shot given a query $Q = \{q_1, q_2, \ldots, q_n\}$ is calculated as a mixture of the language models for shot, scene, video and collection, where scenes are defined as sequences of 5 consecutive shots.

$$\text{score}(\text{shot}|Q) = P(shot) \cdot \prod_{i=1}^{n} [\alpha P(q_i|\text{shot}) + \beta P(q_i|\text{scene}) + \gamma P(q_i|\text{video}) + \delta P(q_i|\text{collection})], \quad (1)$$

where $\alpha + \beta + \gamma + \delta = 1$, and $P(\text{shot})$ is the prior probability of the shot.[1] All experimental runs are compared to a baseline that uses a uniform prior (i.e., $P(\text{shot}) = \frac{1}{N_{\text{shots}}}$, where N_{shots} is the total number of shots in the collection).

5.1 Retrospective experiments

To test whether the surface features could improve retrieval effectiveness if we would have full knowledge of their distributions, we experimented with estimating the distributions from the full relevance judgements. Of course, in a realistic setting, relevance information will never be fully available, but retrospective analysis allows us to explore the potential of the surface features.

The length prior is a linear function of the number of words in the shot's transcript, as is common in the language modelling approach to information retrieval. All other priors are based on the empirical distribution of the surface features in the relevant set. We are interested in the probability of relevance given a surface feature (sf), which can be estimated based on the distribution of the feature in the relevant set and in the collection as follows:

$$P(\text{rel}|\text{sf}) = \frac{P(\text{sf}|\text{rel})P(\text{rel})}{P(\text{sf})} \propto \frac{P(\text{sf}|\text{rel})}{P(\text{sf})} = \frac{\#(\text{rel}, \text{sf})}{\#(\text{sf})}, \tag{2}$$

where $\#(\text{rel}, \text{sf})$ is defined as the number of relevant documents with surface feature sf, and $\#(\text{sf})$ as the total number of documents with that feature. Table 1 lists the priors studied and reports the mean average precision for each.

run name	description	MAP
baseline	no prior	0.075
length prior	size of transcript in words	0.075
source prior	broadcaster (ABC or CNN)	0.081
week prior	week of the year of broadcast	0.087
duration prior	duration of shot in seconds	0.095
minute prior	start of shot in minutes from start of broadcast	0.096

Table 1. Mean average precision (MAP) for various priors estimated on full relevance judgements (retrospective).

The table shows that all studied surface features could potentially improve over the baseline, except for document length. The limited influence of a length prior (especially when compared to its importance in e.g. text retrieval) could be explained from the fact that the shots do not vary much in length. In addition, unlike the other priors, the length prior is not topic-specific. Estimating the length prior on a topic-by-topic basis could perhaps give some improvement,

[1] We use the mixing parameters optimised on the TRECVID2003 test set: $\alpha = 0.4, \beta = 0.4, \gamma = 0.02, \delta = 0.18$

but a large effect is unlikely given the flat distribution of shot lengths, and overfitting is likely.

5.2 Feedback Experiments

After learning that the surface features can potentially improve retrieval results, we experimented in a more realistic setting, one where no full knowledge of the relevant set is available. We took the relevance judgements of the top N results of the baseline run to estimate the surface feature priors from, thus mimicking relevance feedback on the top N retrieved documents. Based on these priors, we produced a new ranking of shots keeping the top N fixed.[2] The new ranking is compared to the baseline using mean average precision. Computing the priors directly using Equation 2 from the small amount of data available in the top N would result in poor estimates due to over-fitting. To avoid this, we interpolate the estimates obtained from the top N with a uniform prior:

$$P(\text{shot}) = \lambda P(\text{rel}|\text{sf}) + (1-\lambda)\frac{1}{N_{\text{shots}}}.$$

This way shots with surface features that are not observed in the relevant documents in the top N still have a chance of being retrieved. We experimented with feedback on the top 20 results ($N = 20$), and with various values of the mixing parameter λ. Table 2 shows the results.

run name	$\lambda = 0.3$	$\lambda = 0.5$	$\lambda = 0.8$	$\lambda = 1.0$
baseline	0.075	0.075	0.075	0.075
source prior	0.068	0.068	0.069	0.073
week prior	0.059	0.059	0.059	0.060
duration prior	0.046	0.046	0.046	0.047
minute prior	0.055	0.055	0.056	0.057

Table 2. Mean average precision (MAP) for priors estimated on Top 20 feedback with different values for the mixing parameter λ.

The mean average precision scores show that, despite the potential, none of the surface feature priors is of practical use in this setup. One of the reasons could be that we are over-fitting on the relevant documents found in the top N. The granularity of measuring some of the features may be too small. Take for example the minute prior, where we look at the starting minute of a shot. Once we have found a relevant shot that starts at a certain minute, we may decide to prefer shots around that time rather than only shots that start at the exact same minute as we have done so far. To investigate this, we take a closer look at the minute prior, and follow the approach of Lin and Hauptmann [6].

[2] Re-ordering the top N is likely to produce a higher MAP, but does not give the user new results.

They use kernel density estimation to estimate the density of specific classes of video (sporting events, weather news) over time. Like Lin and Hauptmann, we use Gaussian kernels, but in our case no classification of the topic is needed; we estimate on a topic-by-topic basis. The width of the kernels (σ) is varied from 1 to 10 minutes. The resulting MAPs are shown in Figure 9. Using these smooth minute priors estimated on either top 10 or top 20 feedback, yields an improvement over the baseline. Taking a closer look at the results reveals that the improvement is due to a few topics only (related to major news events and sports events). Nevertheless, it does not harm the other topics. It is interesting to see that prior information can be obtained from such a small amount of training data, and that it can be used in a practical situation.

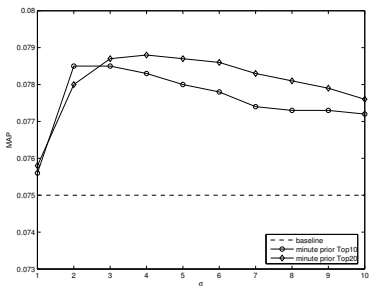

Fig. 9. Mean average precision for minute prior based on Gaussian kernel density estimation.

6 Discussion

This paper has shown that surface features can contain useful information for multimedia retrieval. Knowledge about the relevant features for a given topic can be used to narrow down the search space, or to prioritise documents with certain characteristics. Retrospective experiments with full knowledge of the relevance judgements have shown the potential of using such features. A more detailed study of one of these features, the time of shots within a broadcast, showed that also in a practical situation, surface features can be useful.

The effect of the minute prior may be partially attributed to retrieving the neighbours of known relevant shots. In a general retrieval task, this may not be that interesting, and a sequence of related shots should perhaps be treated as a single retrieval unit. But in the retrieval task studied here, we are searching for shots because of their visual appearance. In such a setting, each shot can be seen as a separate relevant document, since each represents a different view of the same event.

Although the features studied in the present paper and their effectiveness may be restricted to the news broadcast domain, the principle is generally applicable.

As long as a collection is relatively homogeneous, useful surface features are likely to exist.

References

1. Brin, S., Page, L.: The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems **30** (1998) 107–117
2. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. Journal of the ACM **46** (1999) 604–632
3. Lee, H., Smeaton, A.F.: Designing the user interface for the Físchlár digital video library. Journal of Digital Information **2** (2002)
4. Yan, R., Yang, J., Hauptmann, A.G.: Learning query-class dependent weights in automatic video retrieval. In: MULTIMEDIA '04: Proceedings of the 12th annual ACM international conference on Multimedia, ACM Press (2004) 548–555
5. Chua, T.S., Neo, S.Y., Li, K.Y., Wang, G., Shi, R., Zhao, M., Xu, H.: Trecvid 2004 search and feature extraction task by NUS PRIS. In: TREC Video Retrieval Evaluation Online Proceedings. (2004)
6. Lin, W.H., Hauptmann, A.: Modelling timing features in broadcast news video classification. In: Proceedings of the 2004 IEEE International Conference on Multimedia and Expo (ICME), Taipei, Taiwan (2004) 27–30
7. Ponte, J.M., Croft, W.B.: A language modeling approach to information retrieval. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Dev elopment in Information Retrieval. (1998) 275–281
8. Hiemstra, D.: A linguistically motivated probabilistic model of information retrieval. In Nicolaou, C., Stephanidis, C., eds.: Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries (ECDL). Volume 513 of Lecture Notes in Computer Science., Springer-Verlag (1998) 569–584
9. Kraaij, W., Westerveld, T., Hiemstra, D.: The importance of prior probabilities for entry page search. In: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Dev elopment in Information Retrieval, ACM Press (2002) 27–34
10. Ramirez, G., Westerveld, T., de Vries, A.P.: Structural features in content oriented XML retrieval. Technical Report INS-E0508, CWI, Amsterdam, The Netherlands (2005)
11. Gauvain, J.L., Lamel, L., Adda, G.: The LIMSI broadcast news transcription system. Speech Communication **37** (2002) 89–108