From Media Crossing to Media Mining

Franciska de Jong[*†]

[*]University of Twente, Human Media Interaction group
Enschede, The Netherlands
{fdejong}@ewi.utwente.nl
[†]TNO ICT, Delft, The Netherlands

Abstract

This paper reviews how the concept of Media Crossing has contributed to the advancement of the application domain of information access and explores directions for a future research agenda. These will include themes that could help to broaden the scope and to incorporate the concept of medium-crossing in a more general approach that not only uses combinations of medium-specific processing, but that also exploits more abstract medium-independent representations, partly based on the foundational work on statistical language models for information retrieval. Three examples of successful applications of media crossing will be presented, with a focus on the aspects that could be considered a first step towards a generalized form of media mining.

1. Introduction

According to numerous policy makers, IT gurus and advocates of IT initiatives, there is something to be gained by the possibility for users of all kinds to watch or listen audiovisual content anytime, anywhere and via all kinds of platforms. The kind of information referred to by the various kinds of visionaries, can be characterized as heterogeneous in topic, in format, in language, and collection structure. A wide range of reactions on this vision can be observed. From the side of the sceptics one of the major questions is whether the concept of nomadic consumption of arbitrary types of content really match realistic user needs? But there is also the view that the target is right, but that the vision expressed above is not ambitious enough and that the important trend should not focus primarily on the use case of advanced content consumption, but on the more abstract modeling and representation of information, independent of use cases. This dichotomy could be interpreted as two compatible views, corresponding to two different agendas. On the one hand that of software vendors and researchers looking for opportunities to demonstrate the added value of their work, and on the other hand the long term agenda of research communities?

1.1. From Media Crossing ....

With the growing interest in cross-media functionality (for applications such as indexing, browsing, generation, etc.), the insight in the inherent limitations grows as well. Attempts to build cross-media indexing environments have been around now for more than 10 years. Some of these attempts even have yielded unquestionable successes, either commercially or researchwise.

Any serious review of the added value of tools and techniques that have been prosposed and its potential for future advancement of the fiels of information processing will make clear that in the long run we can not be satisfied with a framework that is characterized by the combination of multiple medium-specific processing. Instead, a future should imply a perspective and a vision on how to build a parameterized representation space that could serve both as a foundational framework for formal information models as well as a conceptual basis for application development.

1.2. ... to Media Mining

This paper will describe and compare three different experimental systems that can be considered successful in their contribution to the research agenda. That is, looking backwards. Another way to characterize them is to call them

relatively simple instantiations of media-crossing applications. In what follows below it will be shown how the three experimental systems each call for a more ambitious and multidisciplinary approach that could help define the next-generation of content consumption tools. Though conceptually and formally these tools can clearly be associated with well-establsihed and familiar concepts such as multimedia retrieval or cross-media indexing, a more careful investigation can reveal that they each have taken one or more small steps into a direction that is far more ambitious. For this reason they could be considered to as examples of 'media mining' *avant la lettre.*

In the next section the concept of media mining will briefly be introduced an even widened. The range of research themes that can be linked to it, will be addressed via the description of three cases in the section 3-5. The concluding section will summarize the main findings.

2. Media Mining

Very soon after the introduction of the notion of data-mining in the nineties, it became clear that *'knowledge discovery'*, a term often used for data mining techniques, was not just applicable to the digging up of more or less hidden data patterns in traditional databases. Via text mining, audio mining en media mining, the concept has been claimed to be applicable to all sorts of digital data. Even the term *reality mining* has been introduced and rapidly taken up: Google produced 3000 hits already in Spring 2005, and over 13000 in April 2006). There seems no limit to the applicability of the concept of reality mining: it is hard to come up with something that is NOT covered by the word 'reality'. But even without trying to tackle this philosophical challenge it is easy to see that there is a magnitude of data almost above imagination and that the diversity of types of 'beings' that can be captured with mining techniques is enormous. In principle knowledge about reality and its inhabitants do not follow, let alone obey the borders of media formats and modalities. As a consequence developing tools that support the crossing of format borders in exploring digital archives can be no more than a very first step towards fully exploiting the treasures out there.

Still, and luckily, tools for simply crossing media in search environments do exist. Already for several few years there is even a search task within the context of the TRECVID that stimulates the exploitation of speech transcripts for the retrieval of video. (Cf. (Smeaton et al., 2003).) Relativizing the advancedness of such tools obliges one even more to assess their value. Of course it would be unjust to deny that they can be extremely useful, and particularly in very specific uses cases with very specific user tasks and data sets, crossing media may be the only way to go. But even if one accepts the always-everything-anywhere-mantra, research agenda's should take a wider perspective and develop frameworks that can accommodate more ambitious functionalities for the tackling of the problem of indexing multifaceted collections. Actual user needs for this domain may be hard to predict, but there is clearly an interest from e.g., content syndication parties, portal owners and content providers.

In addition to more and better medium-specific analysis tools, there is the need for analysis models that deliver features that can be integrated in a medium-independent representation and for search models that can abstract away from media-specific features. Ad hoc merging of ranked lists based on word occurrence statistics and image features can be effective, but the real goal should be transformation and integration of representations into one medium-independent representation. The attempt to use conceptual structures as a representation that is independent of language and modality is one of the most salient features of what has more recently become known as *semantic web*, and it will be crucial to take up lessons learned from that framework, including work on the bridging of the *semantic gap*, or more in particular content-based image analysis. And thirdly the metaphor of *translation* could help to clarify the difference in ambition between media crossing and media mining.

As announced, the next sections we will review three experimental approaches each illustrating the media crossing paradigm in a different way.

- Content reduction
- Content merging

- Content enhancement

3. The Content Reduction Case

In various domains, professional information analysts have to deal with large amounts of information which is refreshed on a daily basis and disseminated via various media types: traditional newspapers, news wires and magazines, internet sites and also television broadcasts via air or cable. Analysis and monitoring of these open news sources, which in some cases are coupled to non-public sources of information, is often crucial for efficient and effective workflow. Various mining tools can support the task of news analysts. In this section the crucial role of content reduction as prerequisite for mining will be discussed.

3.1. Parameterized abstraction and summarization

Multimedia news browsing differs from multimedia retrieval in several respects. A major difference is that browsers are supposed to support information search by offering the user not just access to data, but also one or more perspectives on the available data. (Examples of different perspectives are e.g., chronology and geography.) The more flexible a browser, the more different perspectives. etc. In other words: browsers offer a wider range of access functionalities in an integrated way. Indexing can be the basic functionality, but in addition clustering, classification, extraction of headlines and proper names, and summarization can be exploited to build. If a disclosure system integrates news content from heterogenous sources and in multiple formats, e.g., text, audio and video, a salient feature of the browsing functionality could be that that the content can be accessed at various levels of abstraction. For this purpose a variety of content reduction tools can be applied.

Examples of automatic content reduction techniques are abstraction tools such as classification, redundancy detection (via topic-based clustering), summarization, or the generation of a network representation. There is no absolute criterion for the adequacy of these techniques. Whether or not a classification or a summary of a document is useful may vary per user, per user task, per location, etc. To set the system parameters that eventually determine the output, tools that generate abstractions should collect frequency and co-occurrence data of content features, weigh them against background models, and combine them with information about the user and the context of use.

3.2. Abstraction *versus* reduction

Effective content abstraction is a key feature for improved efficiency of the information analysis task. In this context the notion 'abstraction' refers both to conceptual structure, as well as to (reduced) content size. Both forms may pay a role in the automatic enrichment of content via a multifaceted metadata structure.

Various useful levels of abstraction can be distinguished, as different analysis tasks may impose different requirements on the level of conciseness, and even different perspectives on the content can correspond to different metadata requirements. For example, a proper name index on a cluster gives another perspective than a list of topic labels generated by thesaurus-based classification. Metadata types such as keywords and headlines help the user to select potentially interesting clusters for further inspection. This more detailed inspection step can subsequently involve looking at the titles of the individual news items and reading a multi-document extract. Though content abstraction implies content reduction, the reverse only holds if the reduced representations (e.g., summaries, headlines) are representative from one or more perspectives. This is independent of whether the abstraction techniques yield reduced representations in running text, such as extractive or abstractive summaries, or extracted headlines, or structured objects, such as networks of list of proper names or named entities, topic labels for clusters, list extracted key words, etc.

3.3. Lessons learned from Novalist

At TNO a news browser for heterogenous media archives has been developed which is called Novalist. It aims to facilitate the work of information analysts in the following way: (i) related news stories are clustered to create dossiers, sometimes also called 'threads', (ii) dossiers resulting from clustering are analysed and annotated with several types of metadata, and (iii) a browsing screen provides multiple views on the dossiers and their metadata.

One of the reasons why it offers an interesting case for the perspective of this paper is the content base for which the browser functionality can be demonstrated. The corpus disclosed by the demonstrator system consists of a collection of news items published by a number of major Dutch newspapers and magazines, web crawls, a video corpus of several news magazines and a video archive with all 2001 broadcasts of *NOS Journaal*, the daily news show of the Dutch public TV station. The autocue files for the video archive function as collateral text, i.e., text that is not the primary target of search, but that supports the disclosure of video via the time links to media fragments. The entire collection consists of some 160,000 individual news items from 21 different sources.

Another crucial aspect is the technique known as *document clustering* applied in combination with *topic detection* (also known as topic discovery). The system has to deal with dynamic information, about which no full prior knowledge is available. There is no fixed number of target topics and events types. The system must both discover new events as the incoming stories are processed, and associate incoming stories with the event-based story clusters already created. Clustering is done incrementally: for a new incoming story, the system has to decide instantaneously to which topic cluster the story belongs. Since the clustering algorithms are unsupervised, no training data is needed.

Via document clustering, structure is generated in news streams, while the annotations can be applied as filters: search for relevant items need not to apply on analyzed data but can be limited to relevant subsets of the collection. Novalist supports the fast identification of relevant dossiers during browsing. Dossiers are visualized in a compact overview window with links to a time axis. Additional functionality could consist of the automatic generation of links to related sources, both internal and external.

The screen dump of the end-user application in Figure 1 illustrates the browser functionality. For a detailed explanation of the concept of topic detection and the similarity concept applied in the language modeling approach that is underlying Novalist, and for an overview of the performance evaluation of some components, cf. (Spitters and Kraaij, 2002), (Jong and Kraaij, 2005).

Novalist demonstrates that multiple document abstractions effectively mediate different levels of granularity. The analysis can be performed independently of end-user queries. Due to the emphasis on content preprocessing it can support an entire chain of users: content portals that select subsets of news according to filters to serve their users, professional information analysts that link the portal content to their own repositories, and nomadic news consumers. The media crossing proper is limited to linking audiovisual data via their textual transcriptions to the items in the text-based clusters. The source crossing that comes in via the combination of a wide range of open source titles could be viewed as a distinguishing feature as well. But it is the clustering that brings in the mining perspective, strengthened by the fact that the clustering could easily be extended to numerical data, click patterns to set profiles, etc.

Similar dossier generation applications, with topic clustering as basis and content reduction as additional functionality, could be applied in other domains than news, and/or for other combinations of media. In addition to text from newspapers and autocue files (=teleprompter) files, transcripts of broadcast audio generated with automatic speech recognition (ASR) could be taken into account, Assuming that the material can be properly segmented, such sources could be linked to the related topical clusters. Cf. also section 5.

4. The Content Merging Case

In the spectrum of attempts to exploit textual resources for the disclosure of media archives most attention has gone into the role of speech transcripts and their added value on the retrieval of news video broadcasts. Less attention has been given to the possibility to apply information extraction techniques for video archives. The project MUMIS which was completed in 2003 can be considered to fill this gap. In addition to investigating the possibility for the generation of speech transcripts for sports programmes, it payed attention to the possibility to exploit the redundancy in the target document collection. The system components developed provide an analysis for news, commentaries, structured tables from reports, covering international football

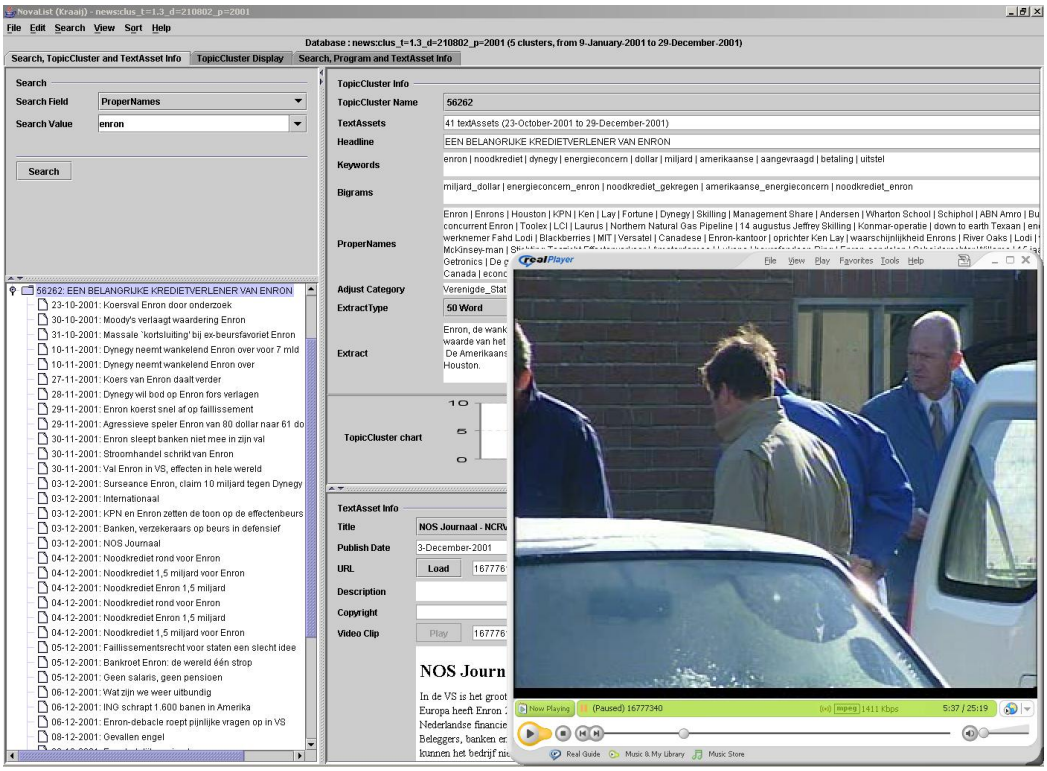

Figure 1: Novalist: browsing multimedia dossiers through associated metadata; query term: 'Enron'

games in multiple languages and multiple modalities, and the resultant data are merged to function as a searchable conceptual knowledge base of all content with links to the timecodes of the corresponding media fragments.

4.1. Multi-source Information Extraction

The combination of research issues central to MUMIS were given in by the characteristic features of the archive studied: football commentaries for an international tournament. What are the crucial features for this type of archive: it consists of video recordings accompanied by several textual sources that cover the same event, but do not necessarily give identical or overlapping information about that event. Cf. Table 1 for examples. Rather the documents should be called parallel, or even less weakly: comparable. As a consequence of the availability of such a combination of sources, the relations between the extraction results can be analysed in order to generate one single merged representation. Errors originating from one of the texts can be removed based on information from the other texts, redundancies can be taken out, and furthermore the merged partial knowledge from separate sources provides a more complete and coherent annotation of the material to be disclosed. The MUMIS disclosure approach can be termed *multi-source information extraction*. As IE modules , such as GATE (cf. (Saggion et al., 2002)) and SCHUG (cf. (Declerck, 2002)), have been applied developed for three languages, it is also a case of *multilingual* IE.

Formal text
England 1 - 0 Germany Shearer (52) Bookings Beckham (42)...
Ticker
41 mins: Beckham is shown a yellow card for retaliating on Ulf Kirsten seconds after he is denied a free-kick.
40' Hoekschop Engeland met David Beckham. Slecht getrapt. Meteen maakt Beckham daarna een fout en krijgt een gele kaart.
Match
David Beckham - a muted force in attack - was shown a yellow card for a late challenge on Kirsten...
Transcription
...it's gonna be a card here for David Beckham it is yellow mmm well again his was the name in the post match headlines...
David Beckham hielt die Sohle noch druber schauen Sie mit dem Hinterteil auch harter Einsatz gegen Kirsten und Collina zeigt ihm Gelb eine der Unarten leider von David Beckham
Beckham met*x Kirsten dat is nou weer dom wat die Beckham doet ja zal ie dat dan nooit leren Kirsten overdrijft nu hoor maar Kirsten gaat 't duel in geeft een zet en dan reageert Beckham op deze manier in ieder geval krijgt ie dan weer geel

Table 1: Different accounts of the same event in different languages

4.2. Improved retrieval via merging

In MUMIS the goal of merging is to yield improved metadata based on information from all documents available from the various sources. As is to be expected, complete recognition of events in natural language sentences is extremely difficult. Often, events will be only partially recognised. The result of merging is one description for all events of a single match that in terms of completeness and correctness has been optimized. A merged annotation is supposed to offer better retrieval results for the multimedia content. The example below, taken from actual results on the Euro 2000 match Netherlands vs. Yugoslavia, gives a rough indication of how merging results in a more complete picture of what happened in the 30-31st minute of the match.

> The IE component recognizes in document A a description of an action of the type SAVE, performed in the 31st minute. In addition, it recognizes the names of two instances of the concept PLAYER: Van der Sar (the Dutch goalkeeper) and Mihajlovic (a Yugoslavian player), but the IE system can not figure out which of these two performed the save.
> In document B IE component recognizes an event of the type FREE-KICK in the 30th minute, and the names of the same two players. It fails to detect which player took the free-kick.
> The fact that the same two players are involved, plus the small difference between the time-stamps, strongly suggests that both descriptions are about the same event. The merger component matches the partial descriptions from A and B, and concludes that it was Mihajlovic who took the free-kick which was followed by a save by Van der Sar.

Figure 2: MUMIS: Example of event merging (informal)

The merging procedure exploits the fact that all available information sources make reference to a time line for the soccer match. This timeline can either be explicit, but sometimes remains implicit. The examples indicates that merging is a combination of three subtasks: time-alignment, unification, and re-ordering. Figure 3 shows the first step for a set of two documents.

As reported in e.g., (Kuper and et al, 2003) experiments have indicated that merging seems to improve retrieval performance.

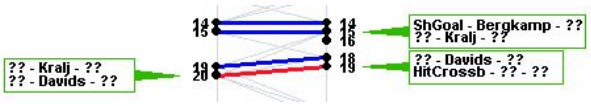

Figure 3: MUMIS: Time-alignment. Vertical lines denote documents, numbers are time stamps; thin lines: possible bindings; thick lines: strongest bindings.

4.3. Lessons learned

MUMIS is clearly a case of information access via media crossing: via time-coded text links into video fragments can be provided. But from the mining perspective it is the merging aspect that is more salient. MUMIS showed that when multiple sources of information are available at the same time, it is likely that the quality and/or reliability will diverge. In combination with IE technology, domain models that link up to templates sets of an interesting size, and reasoning techniques it becomes possible to single out, or even generate optimal representations. Clearly something can be won by having available more, and ideally certified sources. In the MUMIS case the enhancement of content representations via merging exploited primarily text, but integration with all kinds of gazeteers, and repositories of numerical data sources could have additional impact. Especially the scalability of this type of content processing is an issue for further research.

5. The Content Enhancement Case

Most tools that support the searching and browsing of media content in some way or other deploy the concept of *matching*. A representation of the search query is matched onto a representation of the information that is available. In case the formats for query and content representation differ there is a mismatch to be solved. Speech is a carrier of language and therefor a candidate format for content oriented search. But for simple text-based querying of spoken word archives there is an initial mismatch. Only if the either the query or the content has been converted to the format of the other can a matching algorithm be applied. As explained in e.g., (Goldmann and et al, 2005), in the case of spoken document search there are several ways to create matchable representations. Most common nowadays are ap-

proaches that seek to preprocess the audio signal and to apply automatic speech recognition (ASR) to produce a textual transcript. The transcripts can be the basis for a time-coded index that can support the search for audio fragment. This technique has been widely applied for various languages, and in absence of generally applicable tools for video analysis, speech has become the number one entry to large volumes of video content. A clear and widely reported example of media crossing.

Less widespread is work on the exploitation of textual content to complement the speech transcripts. The first role of text is of course to feed the language models needed to build large vocabulary speech recognition. But in addition there are other roles. One of them is that parallel or comparable texts can help to decrease or even eliminate the word error rate of ASR systems. If a manually produced transcript is available, e.g., the minutes of parliamentary sessions, or subtitles for broadcast data, the two parallel texts could be aligned. The timecodes of the ASR transcripts could then be fed into the manually produced text, which in turn could be used for user feedback during search. Also the particular the out-of-vocabulary rate could be decreased: if a (non-perfect) ASR transcript is used a the basis for a search of related text, and the terms referring to named entities in the most similar texts are fed into the language models, a second run of the ASR could yield improved recognition results.

Finally there is of course also the possibility to use an audio fragment as a query for textual documents. An obvious application domain for this option is, again, news. But it works also in other domains than news, e.g., oral history archives, meeting or lecture recordings, digital story telling, etc. In combination with e.g., manually generated minutes, historical studies, policy plans, etc., ASR can provide a welcome or even required additional perspective on the recorded A/V content. Another option to consider would be link generation to geographical data (maps) or other kinds of repositories with a non-linear structure.

Initial experiments with the exploitation of ASR generated transcripts for the search of related text in the cultural heritage domain has been reported in e.g., (Morang et al., 2005).

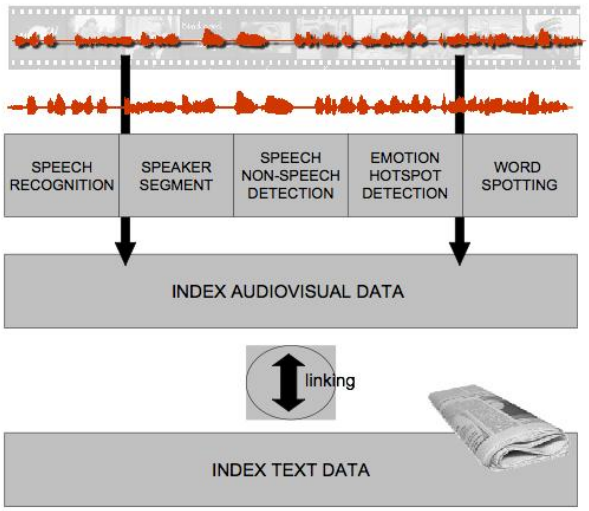

Figure 4: Linking audio to text

6. Conclusions and recommendations

The primary scope of the systems described in this paper are aiming at information access. Pattern detection, which is at the heart of all kinds of mining, is in principle applicable in a much wider range of domains, so the mining concept has a broader scope. Even if we exclude purely numerical data patterns, it could cover diverge topics such as interaction patterns, advertisement campaigns, opinions, moods on certain topics, synonyms, etc. The majority of these examples will serve the needs of professional users first of all (including computational linguists). But as noted above, one of the crucial differences between media mining and media crossing, is the relative independence of use cases for mining. Only after mining tools have been applied the usability of the patterns become evident. In the domain of NLP this holds for example for web-based Question Answering, a search application heavily uses precompiled lists of facts mined from the web.

For development of a research agenda addressing issues that can help building next-generation of media access tools that do not just cross modalities and formats, but that can generate medium-neutral, or in some cases normalized representations, I see a number important sources of inspiration. First there is the analysis of content in term of ontologies, a theme that is getting wide attention under the label of 'semantic web

technology', and that will undoubtedly help to advance the field of media crossing. Second, there is the field of machine translation. Some decades ago it introduced the concepts of interlingual and language-independent meaning representation. The former could be taken as the counterpart of media-crossing approaches that take textual representations as interlingua. The later is close to the generalized interpretation that could help to establish the more general framework and that in this paper has been linked to the concept of media mining. In particular if the number of modalities and formats to be covered increases, the need for medium-independent intermediate representation will increase as well.

Interestingly enough a basis for generalizing approaches to the processing of multimodal information could come from the language (*sic*) modeling approach to Information Retrieval. Cf. (Ponte and Croft, 1998), (Hiemstra, 1998). The language modeling approach to retrieval is based on the philosophy that the language in a relevant document follows the same distribution as that in the query. This same philosophy can also be applied to content-based image and video retrieval, where the only difference lies in the definition of language. Content-based image retrieval systems are usually based on a vector-space model (Smeulders et al., 2000). Collection images are represented as vectors in a high-dimensional feature space, and similarity between images is estimated by a distance metric defined on this space. A drawback of this model is that it is far from obvious how to combine similarity in one representation (e.g., color histograms) with that of another one (e.g., texture); especially when a combination is concerned of different modalities, such as video shots represented by their visual, audio, and speech content. Recently, several attempts have been made to investigate whether discrete bag-of-words models (as used in full-text retrieval) can be also developed and effectively implemented for visual content and the so-called 'visual words' it consists of. Cf. e.g., (Squire et al., 2000), (Jin and Hauptmann., 2002), (Vries and Westerveld, 2004).

With the availability of more abstract models, media formats can not only be crossed, but eventually really integrated, and the content they store can be explored in a genuinely general way. The approaches mentioned rely on heavy processing power, but with the likely advances in grid-processing there is no reason to doubt that the required capacity will become available. While seeking on the one hand collaboration with communities involved in foundational research on abstract content models, language engineers should on the other hand continue to carefully design the media-crossing applications, and to apply all the heuristics they can get hold of to improve their tools and to demonstrate the added value of language processing for the disclosure of information.

Acknowledgements

This work was partly supported by the Dutch bsik-programmes MultimediaN (www.multimedian.nl) and BioRange (www.nbic.nl), and the EU projects AMI (IST-FP6-506811), MESH (IST-FP6-027685), and MediaCampaign (IST-PF6-027413).

7. References

Declerck, T., 2002. A set of tools for integrtaing linguistic and non-linguistic information. In *Proceedings of SAAKM 2002, ECAI-2002*. Lyon.

Goldmann, J. and S. Renals et al, 2005. Accessing the spoken word. *International Journal on Digital Libraries*.

Hiemstra, D., 1998. A linguistically motivated probabilistic model of information retrieval. In *Proceedings of ECDL 1998*.

Jin, R. and A.G. Hauptmann., 2002. Using a probabilistic source model for comparing images. In *Proceedings International Conference on Image Processing (ICIP02)*. Rochester, NY.

Jong, F.M.G. de and W. Kraaij, 2005. Content reduction for cross-media browsing. In H. Saggion and J.-L. Minel (eds.), *RANLP workshop 'Crossing Barriers in Text Summarization Reserach*. Borovets, Bulgaria.

Kuper, J. and H. Saggion et al, 2003. Intelligent multimedia indexing and retrieval through multi-source information extraction and merging. In *18th International Joint Conference of Artificial Intelligence (IJCAI)*. Acapulco, Mexico.

Morang, J., F.M.G. de Jong, R.J.F. Ordelman, and A.J. van Hessen, 2005. Infolink: analysis of dutch broadcast news and cross-media browsing. In *Proceedings of CBMI 2005*. Amsterdam.

Ponte, J. and W. Croft, 1998. A language modeling approach to information retrieval. In *Proceedings of the 21st ACM Conference on Research and Development in Information Retrieval (SIGIR'98)*.

Saggion, H., H. Cunningham, D. Maynard, K. Bontcheva, O. Hamza, C. Ursu, and Y. Wilks, 2002. Extracting information for automatic indexing of multimedia material. In *Proceedings of LREC 2002*.

Smeaton, A.F, W. Kraaij, and P. Over, 2003. Trecvid - an overview. In *Proceedings of TRECVID 2003*. USA: NIST.

Smeulders, A.W.M., M. Worring, S. Santini, and R. Jain A. Gupta, 2000. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380.

Spitters, M. and W. Kraaij, 2002. Unsupervised clustering in multilingual news streams. *Proceedings of the LREC 2002 workshop: Event Modelling for Multilingual Document Linking*:42–46.

Squire, D. McG., W. Muller, H. Muller, and T. Pun, 2000. Content-based query of image databases: inspirations from text retrieval. In *Pattern Recognition Letters*, volume 21.

Vries, A. de and T. Westerveld, 2004. A comparison of continuous vs. discrete image models for probabilistic image and video retrieval. In *Proceedings International Conference on Image Processing (ICIP'04)*.