The CHORUS gap analysis on User-centered methodology for design and evaluation of multi-media information access systems

Jussi Karlgren
SICS, Stockholm
CHORUS project
http://ist-chorus.org

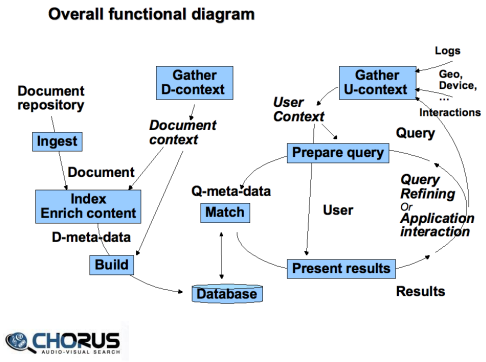

Figure 1. The CHORUS functional description.

The CHORUS Coordination action

CHORUS is a Coordination Action, a specific type of project funded by the European commission under its research programmes, intended to bring together research projects with common goals, in the field of search technologies for digital audio-visual content, one of the strategic objectives of the current research frame program. CHORUS coordinates a number of research projects in the general area of audio-visual and multi-media information access and management.

The most important single contribution of the CHORUS work plan will be to provide a survey of the field and a roadmap with a gap analysis for the realisation of viable audio-visual search engines by European partners. This is done by several means. CHORUS organises Think-Tanks with industrial participation, focussed workshops to treat specific questions, and more general conferences for academic discussions. CHORUS is now in its final phase, and is currently preparing its final report together with a final conference to mark its publication.

As part of this work CHORUS will establish a reference functional description of multimedia search engines. A technology view of the functional description is given in Figure 1. This functional description will be related to use cases and scenarios gathered from interviews with active research projects and industrial activities.

Challenges, Bottlenecks, Thresholds

The CHORUS conferences, think-tanks, and workshops have collated major challenges to providing useful multi-media information access services as identified by academic and commercial projects. A tentative generalisation is given by the following points.

Knowledge representation

Multi-medial information objects are not as simply segmentable into meaningful micro-items as are texts. Enabling search and indexing is a greater challenge for non-textual items. Several projects use existing proxies for information such as user-contributed comments or tags; others use contextual data mined from user behaviour or meta-data harvested from extraneous knowledge sources. This invites a debate as to whether data in different media are of inherently different type: are textual data primary and more refined where e.g. video data are more raw, in some respect? Can we envision a general representation not tailored to some task or domain? Should we aim to build e.g. a language-based representation for data which haven't been expressed in linguistic terms — or is text a false target for representation? Can we presume to know what the tentative uses of some information object will be when we are working with media where the space of possible application broadens and the leeway the data affords for individual interpretation is larger?

Search Strategies and Session Design

The interaction point between user and system can be designed to conform to various usage scenarios: browsing, searching, exploring, formulating recommendations, receiving recommenda-

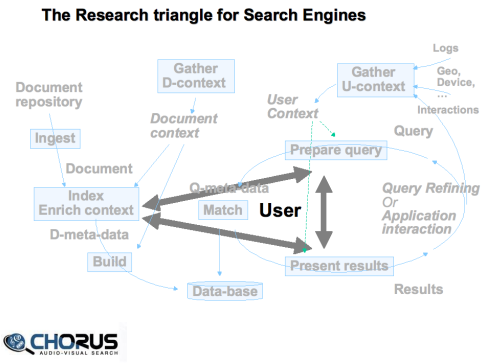

Figure 2. Research questions related to usage are distributed over the functional description.

tions — and naturally, this is not an exhaustive list. Is this a modular issue, not related to underlying system architecture, or does the choice of interaction paradigm have ramifications for engineering the system itself? And does the character of the data, viz. multi-medial data in the case at hand, influence or incur preferences for the choice of possible session designs?

User-contributed data vs Editorial provision of data

The lowered publication threshold and true, if not completely symmetric, bi-directional communication technology allows users to contribute content, and structure it at will and on demand. A recurring theme in recent projects presented was how to provide a framework which a) encourages and motivates users to contribute to shared information systems and b) provides guidance, quality assurance, and a shared semantic space to contributions, better to build a common body of knowledge. Much of the user-generated content will be used the same way that professional content is, but the models for data quality, persistence, archival character, intellectual rights issues differ importantly.

Impact

On an overview level, the question of what characterises success for an multimedia information access project is at the forefront of academic and commercial projects and research funding agencies. How can a multimedia information access project achieve impact? Is the technology, the content or the service design the major determining factor? Is it all a question of marketing?

Taking the various challenges together brings the problem of evaluation to the fore. Evaluating information retrieval has traditionally been done using the target notion of topical relevance. How this target for quantitative relevance can be extended to non-topical access scenarios is not obvious – should it be supplanted by more general notions such as user satisfaction or pertinence or should the notion of relevance be enhanced?

Topical evaluation schemes in information retrieval

Some types of data are inherently more topical than others. For factual narrative and topical argumentative texts and text-like materials, the field of information retrieval provides a framework for organisation, systematic categorisation, search, and retrieval of items or sets of items. There are established use cases, success metrics, and evaluation benchmarks of various kinds, which all based on the target notion of topical *relevance*. Much of the success of information retrieval as a research field is owed to its handy and formal evaluation schemes.

Character of interaction; character of target data

The evaluation framework from information retrieval research does not carry over well to the study of non-textual information access. Returning to the challenges given above, several of them have direct pertinence for evaluation. Interaction with information access systems is not necessarily limited to users entering one brief query, as is typical for topical text retrieval systems.

Extending information retrieval to other media for topical retrieval is conceptually, if not always practically, straight-forward, but moving retrieval to cater for other facets of usefulness, pertinence or acceptability is more challenging than it first might seem. Firstly, the rationale of users may be different. For usage which is more directed towards less urgent entertainment rather than fulfilling timely information needs, user satisfaction is less obviously modellable. This is an argument for more wide-ranging target metrics, beyond topical relevance.

Secondly, the media itself and their content are a factor against topical evaluation. While some video material may by topically analysable: newscasts, instruction clips, or lectures, e.g. others are intended to provoke a sensation or provide momentary enjoyment – finding common content features over a set of such materials is not obviously possible before the fact. It can be argued that most important factual content is in linguistic form (and thus amenable to topical indexing) and other materials should be accessed in completely different ways. This calls into question the primacy of topical indexing: the functional description given

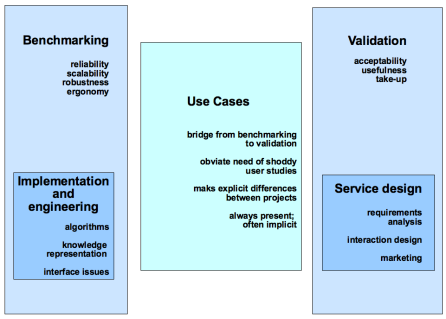

Figure 3. Use cases bridge benchmarking and validation.

by CHORUS does not presuppose topic as the only or even primary indexing criterion.

Thirdly, the interaction with multi-media information systems are more diverse than those to typical text-based systems. If the interaction is based on system-initiated recommendation rather than user-initiated search, on likeness to examples, rather than goal-directed matching to queries, on satisfaction rather than optimisation, the target metrics must be different. Ranking the output of the information system, and the evaluation of that ranking must be done using different schemes than today.

Evaluation beyond benchmarking

The formalisation of evaluation metrics and the *quantitative and comparative methodology* adopted by the information access research field are an important step which sets information access on firmer methodological basis than other related fields in computation and language technology.

To move from abstract benchmarking to more user-sensitive evaluation schemes, the CHORUS Coordination action is in the process of formulating a comprehensive typology of *use cases* for multimedia information access. This will enable development and research efforts to identify likenesses between projects, to leverage previous knowledge, to avoid retreading previous erroneous tracks. It will afford a possibility to parametrize system characteristics visavi features of typical usage, without necessarily performing laborious user studies to do so. It also makes explicit the two often conflated goals of system evaluation: that of system qualities as an engineering effort and that of system usefulness for particular tasks as indicated in Figure 3.

Anchoring the use cases in usage and practice, and predicting user take-up and project impact will need a third methodological step. Beyond a methodology of quantitative and comparative benchmarking, and beyond an understanding of use cases the field needs *methodology for user studies* to establish the effect of technology and a level of craftmanship as regards system design to reduce noise.

Lack of overlap

Human-machine interaction is a huge research field in its own right, and interaction design is a trade with competence, craftmanship and established success criteria. Hitherto, multi-media information access projects relatively seldom have identified interaction as a pressing issue – technology and system factors have overridden those concerns. Simultaneously, the human-machine interaction research field relatively seldom uses multi-media information access as domain of study, nor are interaction design professionals employed to provide a solid basis for user-oriented experimentation.

Generalisable results and guidelines on interaction in multi-media information access need appropriate methodology and craft from the interaction field.

Conclusion

Evaluation of multimedia systems needs to be sensitive to all the specific challenges in play: content representation, use case and session design, and further lowered publication thresholds. Evaluation schemes have been the backbone of text retrieval research for the past decades: the field of multimedia retrieval should take care not to diverge from that tradition for reasons of convenience only – the target concept of relevance may not need to be replaced, but will need either extension or deconstruction to carry over to new use cases, new scenarios, and new types of media. Some of the basic tenets of information access evaluation may need to be modified or given less importance: reproducibility, e.g., is likely to be interpretable not to require a unchangeable data set but an unchangeable situation.

The CHORUS project will present a gap analysis and a roadmap to point out directions to which evaluation and design of future information access systems should direct their attention. This presentation will present this roadmap, currently being authored by the CHORUS working group.