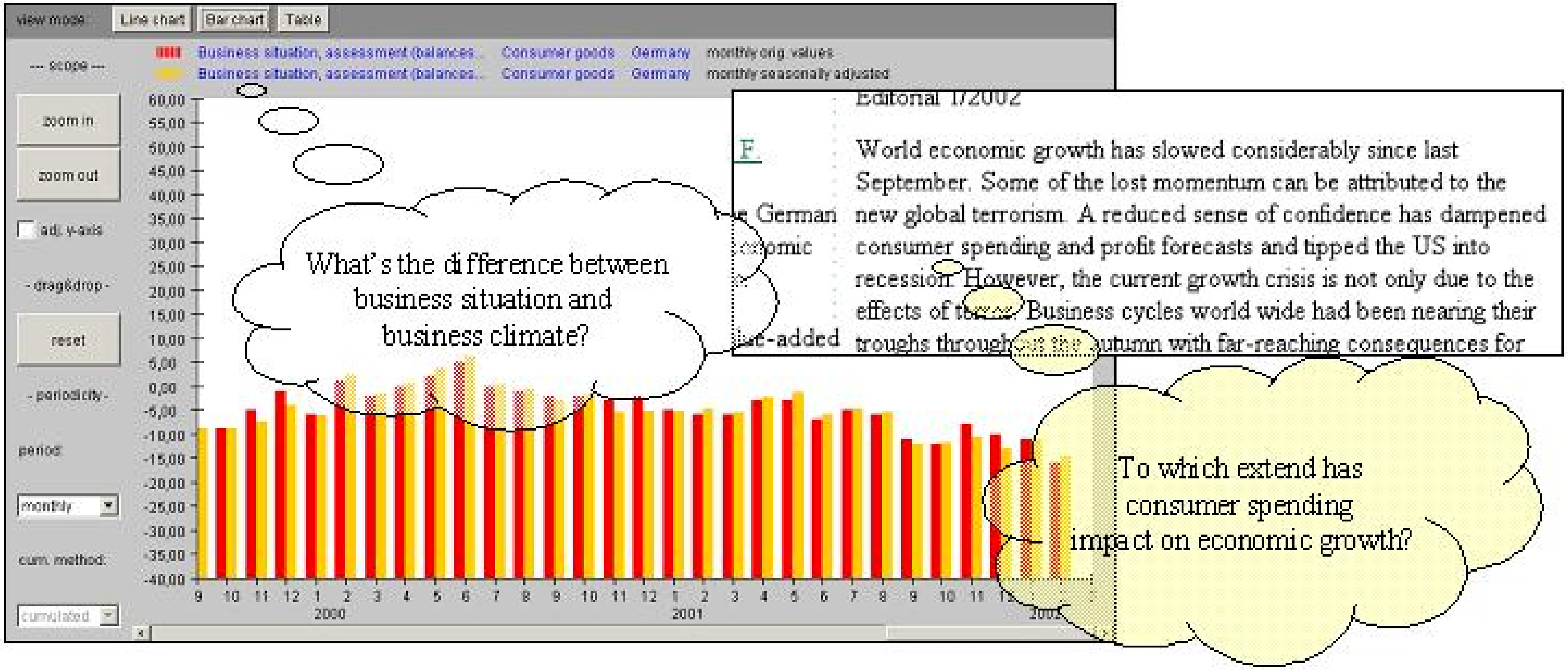
A learning companion - design of personal assistance in an adaptive information and learning ambience

Kurt Englmeier, German Institute for Economic Research (DIW), Berlin, Germany,
Josiane Mothe, Université Paul Sabatier, Institut de Recherche en Informatique de Toulouse, France,
Marios Angelides, Department of Information Systems and Computing, Brunel University, London, UK.

The approach of a learning companion as presented here bases on experiences gained during design, implementation, and user evaluation of IRAIA an information provider platform technology for economic information (IRAIA, IST-1999-10602) and the further application of the technology for a box of inspirations for European puppetry (eStage, IST-2000-28314).

1 Learning features in an information retrieval environment

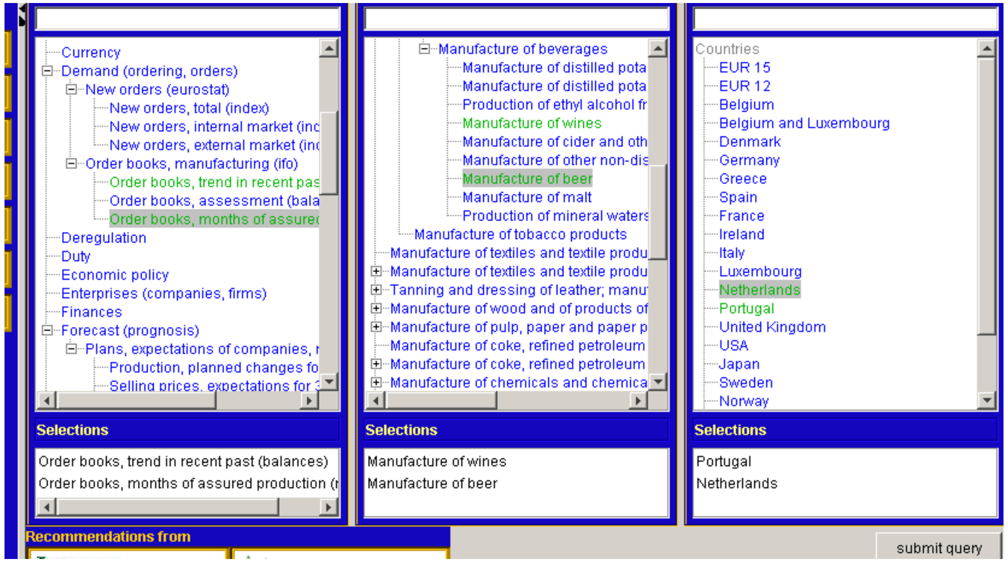
Traditional text- and guidebooks carefully guide users through the topics of a course or a content domain in a specific order the authors have deemed most appropriate. Integrated into a retrieval environment they can be pretty helpful to guide the users to explanations or more details on a specific topic. They can broaden a certain topic or even compensate for missing knowledge by adding explanations to the pages a reader visits. If the user, for instance, is not so familiar with the concept of "business climate" used as an indicator measuring business activities, she or he can switch from the retrieval section to a document or a set of documents that explains this concept, and come back afterwards to continue retrieval.



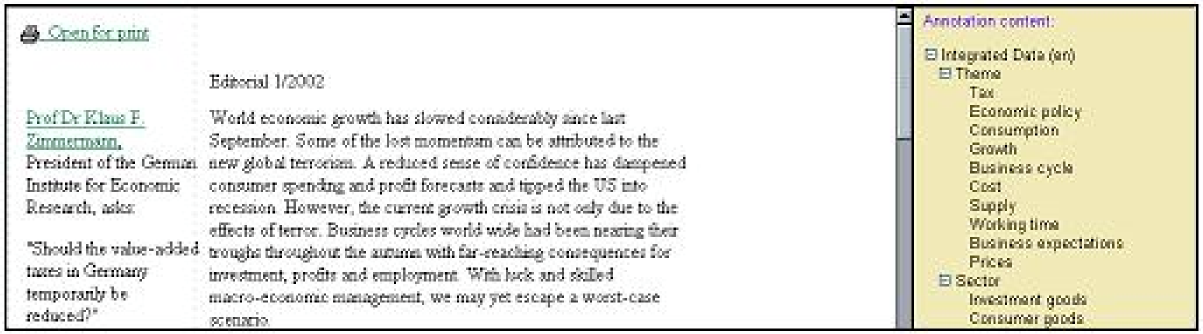
2 Situation-sensitive information spaces

The design of a situation-sensitive information ambience starts with a concise, coherent, and comprehensive abstraction of content that in the end can also reflect a particular retrieval situation. The abstraction of content arises from data when they are combined, arranged, and presented accordingly on a suitable level of abstraction. Only an appropriate combination of texts, images, graphs and the like is in the position to convey the information that is contained in these usually separate and otherwise not fully perceptible components.

2.1 Semantic co-ordinate system



Searching and navigation in a semantic coordinate system. Selected concepts make up the initial query profile. While realising his retrieval strategy the user usually performs iterative steps of defining a query and analysing the retrieved results. In IRAIA, the documents are annotated solely with entries from the hierarchies.



While viewing a document (a text or time series) the interface shows also the annotated concepts (similar to the screen above) that are thus already familiar from the initial query formulation. Modifying the set of annotated concepts is thus tantamount with repetitive query formulation.
By clicking on a certain concept or a brief sequence of concepts the user can evoke a personal assistant that looks after explaining material suitable in the actual situation.

2.2 Inspirational desire

For the ambience of cultural information related to puppetry IRAIA provides an adaptation opportunity that brings serendipity into focus. Overt expressions of serendipity combined with practicality are rare in the realm of information retrieval engineering. The adaptability designed for the query feature in IRAIA can achieve the marriage of technical solutions and inspirational desire. It meets a collection's structural requirements while evoking a sense for exploring an immediate information ambience. Its flamboyance inextricably weds the stunning and the practical.

The goal here is to be a source of inspirations for the ideas of puppeteers – including professionals and amateurs. The things puppeteers are after can be best clarified by the following example from a talk with the ensemble of a puppet theatre: "We were looking for a new play for children that should have some educational effect through a certain level of cruelty as it occurs likewise in many of Grimm's fairy tales. We looked for children's books in libraries, bookshops and even toy shops and run into a book in English telling the story of a mother and cannibal looking for a child to eat. The design idea for puppet came from a picture in a festival announcement showing a puppet representing a fierce devil."

An example of results retrieved by an information portal for puppetry. The portal bases on IRAIA technology, but has more emphasis on serendipity.

Exact retrieval results are marked as "Best results" in the workspace, while the others can be located more in the semantic surroundings of the best match.

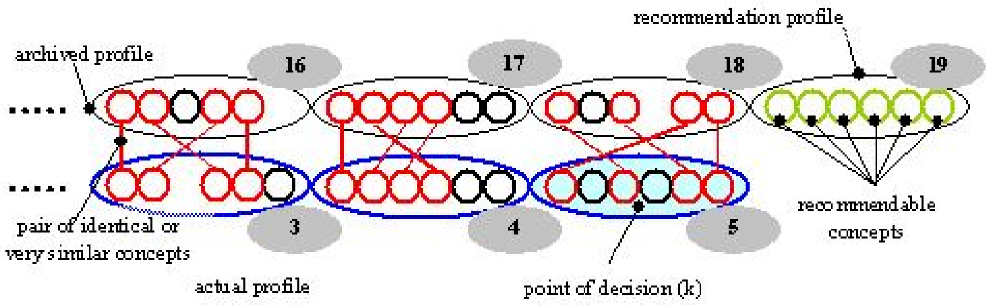
The cornerstones of this user interactions are the following:

- The access to a digital archive must enable the creative talents to go directly to the spot they are looking for while bypassing the vast majority of the collection.
- The interaction mode must cope with the requirement that users show up with a very vague idea that steers their searching and navigating. However, this idea is the only steering element.
- The structure of guidance provided by the system must be sensitive enough to let the users keep their vague idea as the steering element. Any too rigid structure may cause an unacceptable information overload.
- There is no such thing like a precise answer to this vague query. Anything that relates to it can be useful. Thus, any kind of information found by serendipity is welcomed.

3 Personal assistance: the learning companion on stage

If an information space is arranged along a semantic coordinate system each information item can be positioned at its most adequate point within this space. The corresponding entries of the coordinate system reflect the item's annotated concepts. Tasks like satisfying an information need as well as explaining a certain phenomenon in economy require usually a certain set of information items. Such a task-specific information set can be represented as a trace of nodes within the information space. This approach can easily be extended to adaptive recommendations for learning purposes. If the system recognises, for instance, a user certain sequence of documents (items) along a trace it can extract situation-sensitive aspects from this trace or – strictly speaking from its representation on the abstraction level. This means recommendation traces base primarily on observations of the user's search and navigation strategies.

The personal assistant looks for an item from the stock of learning material that could be recommended as explaining document for the actual situation. (Comparing documents base on the matching process outlined in the following chapter.)
The upper profiles (16 through 19) are archived ones. Due to concept similarities up to the decision point (k) and the availability of a further profile (19) that reaches beyond k they are suitable for recommendation. The lower profiles reflect the actual navigation of the user. Pairs of similar concepts are marked by circles with links in between them. The order of the nodes within a query profile is not important. In this example, the users asks for recommendation while completing query no. 5 and gets concepts from profile 19.



The language model in our approach consists first of all of a controlled vocabulary. For a better "understanding" of the meaning of an economic situation expressed a text passage we combine the controlled vocabulary with a discourse model that helps to identify correctly typical features in economic texts. These features usually describe an important behaviour of an economic item. Examples for such feature information are "raise" in "number of sales raised" or "slow down" in "according to business situation assessment investment activities slow down". The discourse model identifies cue words and syntactic structures that function as discourse markers. Like taxonomy concepts themselves the discourse markers are classified as part of "decline", "increase" etc. This information contained in the already annotated concepts is thus enriched by such discourse labels. In an additional step the model can use this information to develop a more generic discourse structures appearing in the presentation of economic phenomena. Moreover, these generic structures support the automatic compilation of information items, a feature useful in information broadcasting.

4 The matching process

In the end, every design model for an information providing system goes back to the good old quest to find an efficient matching between the representation of documents and the user's information need. The querying and navigation process itself can be regarded as a probabilistic inference process that compares document representations (descriptors) based on different forms of linguistic and/or statistical evidence on the one with representations of information needs (queries) based on similar evidence on the other side. The framework for the design of the matching processes is generally known as the inference network model.

The task to be carried out corresponds to document categorisation. While parsing the concept hierarchies the entries of their nodes are treated as profiles. These profiles are composed of phrases because the entries of the concept hierarchies usually consist of a couple of words describing a concept. From the lexical analysis results an index list of words that is in this form a parsed representation of the controlled vocabulary. Alternatively, a profile can be composed of a set of phrases automatically extracted during a learning stage. This latter method is used when a training document set is available. Thus, a node together with its ancestors can be regarded as a pseudo-document that may be helpful in the further matching process. This process compares, roughly speaking, the content of the concept hierarchies with those of the documents and decides which tree nodes are the most prominent ones for this document.

The following evaluation calculates the belief in a profile (a node in a concept hierarchy) due to the occurrence of a concept c_i of a document d:
where

$$bel_{Eh}(d) = \sum_{c_i} \left(\frac{tf_i}{s(d)} \cdot \frac{s(h)}{tfE_i} \right) \cdot e^{\frac{s(Eh \cap d)}{s(Eh)}}$$

Eh= an entry or node from the concept hierarchy h, represented by a profile, C_i = a concept from the profile Eh, tf_i = frequency of the concept C_i in the document d,
tfE_i = frequency of the concept C_i in the node Eh, s(d) = size of the document d (number of concepts), s(h)= size of the concept hierarchy, s(Eh) = size of the node Eh, $s(Eh \cap d)$ = number of terms from Eh that occur in Eh.

$\frac{tf_i}{s(d)}$ *measures the importance of the concept in the document,*

$\frac{s(h)}{tfE_i}$ *measures the importance of the concept in the concept hierarchy and* $\frac{s(Eh \cap d)}{s(Eh)}$ *measures the rate of occurrence of the concepts from the concept hierarchy node in the document (coverage).*

This formula is derived from the well-known *tf.idf* weighting function. The values are normalized and remain between 0 and 1. Concept phrases for a given document are ranked by this function. A concept phrase is selected for annotation if its value is above a certain significance threshold (let's say 0.4, for instance). Due to practical reasons only the three entries highest in ranking as well as above the threshold are chosen.

$$ann(d) = \bigcup_h \begin{cases} bel_{Eh}(d) \text{ with } bel_{Eh}(d) >= \delta_0 \\ \{\} \text{ otherwise} \end{cases}$$

where

δ_0 *= significance threshold, h = concept hierarchy*

Calculations are performed for all concept hierarchies. In principle, this evaluation can be applied at the entire document or parts of it as titles, paragraphs, and the like. We emphasize once again that the matching process bases exclusively on concept phrases.