A Probabilistic Framework for Information Modelling and Retrieval Based on User Annotations on Digital Objects

Ingo Frommholz
University of Duisburg-Essen*
ingo@is.inf.uni-due.de
http://www.is.inf.uni-due.de/staff/ingo.html.en

Annotations are a means to make critical remarks, to explain and comment things, to add notes and give opinions, and to relate objects. Nowadays, they can be found in digital libraries and collaboratories, for example as a building block for scientific discussion on the one hand or as private notes on the other hand. We further find them in product reviews, scientific databases and many "Web 2.0" applications; even well-established concepts like emails can be regarded as annotations in a certain sense. Digital annotations can be (textual) comments, markings (i.e. highlighted parts) and references to other documents or document parts. Since annotations convey information which is potentially important to satisfy a user's information need, this thesis tries to answer the question of how to exploit annotations for information retrieval. It introduces POLAR, a probabilistic, logic-based framework for annotation-based retrieval, and gives a first answer to the question if retrieval effectiveness can be improved with annotations.

A survey of the "annotation universe" reveals some facets of annotations; for example, they can be content level annotations (extending the content of the annotated object) or meta level ones (saying something about the annotated object). Furthermore, there might be positive or negative annotations. Besides the annotations themselves, other objects created during the process of annotation can be interesting for retrieval, these being the annotated fragments. These objects are integrated into an object-oriented model comprising digital objects such as structured documents and annotations as well as fragments. In this model, the different relationships among the various objects are reflected. From this model, the basic data structure for annotation-based retrieval, the structured annotation hypertext, is derived.

In order to thoroughly exploit the information contained in structured annotation hypertexts, a probabilistic, object-oriented logical framework called POLAR is developed. In POLAR, structured annotation hypertexts can be modelled by means of probabilistic propositions and four-valued logics. POLAR allows for specifying several relationships among annotations and annotated (sub)parts or fragments. Queries can be posed to extract the knowledge contained in structured annotation hypertexts. POLAR supports annotation-based retrieval, i.e. document and discussion search, by applying an augmentation strategy (knowledge augmentation, propagating propositions

*Now at Department of Computing Science, University of Glasgow. Email: ingo@dcs.gla.ac.uk.

from subcontexts like annotations, or relevance augmentation, where retrieval status values are propagated) in conjunction with probabilistic inference, where $P(d \rightarrow q)$, the probability that a document d implies a query q, is estimated. POLAR's semantic is based on possible worlds and accessibility relations. It is implemented on top of four-valued probabilistic Datalog, which allows for the reflection of the polarity of annotations and for the consideration of possible inconsistent knowledge in discussion threads.

POLAR's core retrieval functionality, knowledge augmentation with probabilistic inference, is evaluated for discussion and document search. The experiments show that all relevant POLAR objects (annotation targets, fragments and content annotations) are able to increase retrieval effectiveness when used as a context for discussion or document search. Additional experiments reveal that we can determine the polarity of annotations with an accuracy of around 80% by applying machine learning techniques.

Available online at `http://duepublico.uni-due.de/servlets/DerivateServlet/Derivate-21216`