Efficient Crowdsourcing for Metadata Generation

Wolf-Tilo Balke
IFIS, TU Braunschweig,
Braunschweig, Germany
balke@ifis.cs.tu-bs.de

Abstract

Rich and correct metadata still plays a central role in accessing data sources in a semantic fashion. However, at the time of content creation it is often virtually impossible to foresee all possible uses of content and to provide all interesting index terms or categorizations. Therefore semantic retrieval techniques have to provide ways of allowing access to data via missing metadata, which is only created when needed, i.e. at query time. Since the creation of most such metadata will to some degree depend on human judgement (either how to create it in a meaningful way or by actually providing it), crowdsourcing techniques have recently raised attention.

By incorporating human workers into the query execution process crowd-enabled databases already can facilitate intelligent, social capabilities like completing missing data at query time or performing cognitive operators. Typical examples are ranking tasks, evaluating the correctness of automatically extracted information, or judging the similarity or subjective appeal of images. But for really creating metadata for probably large data sources, the number of crowd-sourced mini-tasks to fill in missing metadata values may often be prohibitively large and the resulting data quality is doubtful. Instead of simple crowd-sourcing to obtain all values individually, in this talk utilizing user-generated data found in the Social Web is discussed

By exploiting user ratings semantically meaningful perceptual spaces can be built, i.e. highly-compressed representations of opinions, impressions, and perceptions of large numbers of users. Then, using few training samples obtained by expert crowd sourcing, missing metadata can be extracted automatically from the perceptual space with high quality and at low costs. First experiments show that this approach actually can boost both performance and quality of crowd-enabled databases, while also providing the flexibility to expand schemas in a query-driven fashion.

(Dagstuhl Seminar Series Nr 12171. 22. - 27.04.12
G. Antoniou, O. Corcho, K. Aberer, E. Simperl, R. Studer:
Semantic Data Management)