From People to Entities: Typed Search in the Enterprise and the Web

Gianluca Demartini
L3S Research Center, Hannover, Germany
demartini@L3S.de

Abstract

The exponential growth of digital information available in Enterprises and on the Web creates the need for search tools that can respond to the most sophisticated informational needs. Retrieving relevant documents is not enough anymore and finding entities rather than textual resources provides great support to the user both on the Web and in Enterprises. Many user tasks would be simplified if Search Engines would support typed search, and return entities instead of just Web pages. For example, an executive who tries to solve a problem needs to find people in the company who are knowledgeable about a certain topic. Aggregation of information spread over different documents is a key aspect in this process.

Finding experts is a problem mostly considered in the Enterprise setting where teams for new projects need to be built and problems need to be solved by the right persons. In the first part of the thesis, we propose a model for expert finding based on the well consolidated vector space model for Information Retrieval and investigate its effectiveness.

We can define Entity Retrieval by generalizing the expert finding problem to any entity. In Entity Retrieval the goal is to rank entities according to their relevance to a query (e.g., "Countries where I can pay in Euro"); the set of entities to be ranked is assumed to be loosely defined by a generic category, given in the query itself (e.g., countries), or by some example entities (e.g., Italy, Germany, France). In the second part of the thesis, we investigate different methods based on Semantic Web and Natural Language Processing techniques for solving these tasks both in Wikipedia and, generally, on the Web. Evaluation is a critical aspect of Information Retrieval. We contributed to the field of Information Retrieval evaluation by organizing an evaluation initiative for Entity Retrieval.

Opinions and other relevant information about entities can be provided by different sources in different contexts. News articles report about events where entities are involved. In such setting the temporal dimension is critical as news stories develop over time and new entities appear in the story and others are not relevant anymore. In the third part of this thesis, we study the problem of Entity Retrieval for news applications and the importance of the news trail history (i.e., past related articles) to determine the relevant entities in current articles. We also study opinion evolution about entities. In the last years, the blogosphere has become a vital part of the Web, covering a variety of different points of view and opinions on political and event-related topics such as immigration, election campaigns, or economic developments. We propose a method for automatically extracting public opinion about specific entities from the blogosphere.

Available online at `http://www.gianlucademartini.net/research/phd/`