Proposal for a Master-Thesis

Determination of Document Similarity based on Bayesian Statistics for a Big-Data Information Retrieval Solution

1 Background

dibuco GmbH develops an information retrieval middleware solution for big-data scale corporate data collections (data lake approach). This solution is based on a novel concept of multi-dimensional structuring of a data lake using various notions of data similarity. These can be business driven similarities that express similarities based on the business semantics of the data. But also, business agnostic similarities like textual, language semantics or structural similarities can be used to create views of the data lake related to specific use cases of information discovery and retrieval.

A stateful query API supports such information discovery and retrieval and in particular exploratory approaches. The targeted amounts of corporate data of very high variance in source, size, business purpose and data format do not allow to query the data lake using traditional keyword or phrase searches. We expect that the end-user will not have the necessary knowledge about the information available to be able to express focused keyword queries that result in the relevant available information. Instead the solution is using a novel sample based query approach where the user presents known, relevant information at hand and the system will be responding with similar alternative information, ranked by a confidence of relevance. The explorative query approach is supported by allowing to build and transform a scope of interest using similar information and gradually narrowing down the focus of search.

Therefore, it is necessary to determine the similarity of entities such as documents. According to the huge number of documents to be considered scalability is important. There are different approaches to similarity determination, e.g., stochastic approaches and approaches based on Bayesian statistics. In this thesis, the student investigates approaches of similarity determination for documents based on Bayesian statistics and evaluates their scalability.

2 Tasks

The thesis has two parts: a conceptual part that investigates existing approaches and their adaptation to the information retrieval middleware solution and a development and evaluation part where selected algorithms for document similarity detection are prototypically implemented and evaluated concerning their scalability.

The tasks are as follows:

- Analysis and documentation of use cases for document similarity determination covering both similarities based on business semantics and business agnostic similarities.
- Start of the art analysis of existing approaches for document similarity determination based on Bayesian statistics.
- Selection of approaches to enable document similarity determination in the information retrieval middleware solution for big-data scale corporate data collections.
- Documentation and specification of the selected approaches.

Franz-Schubert Str. 75
D-70195 Stuttgart
tel +49 711 699 475 60
fax +49 711 699 475 62

info@dibuco.de
www.dibuco.de
Geschäftsführer
Dr. Joachim Baumann, Albrecht Stäbler

Sitz der Gesellschaft Stuttgart
Handelsregister Amtsgericht Stuttgart
HRB 753495
USt-IdNr. DE 302017247

Bankverbindung BW-Bank
IBAN: DE07 6005 0101 0008 0997 21
BIC-/SWIFT-Code: SOLADEST600

- Prototypical reference implementation of a subset of the selected approaches to cover the most common use cases identified in the first step.
- Evaluation of the scalability of the realized approaches based on selected sample data

3 Required previous knowledge and experiences

- Good knowledge of Java
- Knowledge of Bayesian statistics and determination of document similarities
- Interest in the general information retrieval problem domain
- **… or the declared intention to deeply dive into these topics in advance**
- Good English language skills (Accompanying result documentation need to be in English language)
- Thesis can be written in German or English