Universität des Saarlandes
Max-Planck-Institut für Informatik
AG5

Learning Rules With Numerical Constants in Large Uncertain Knowledge Bases

Masterarbeit im Fach Informatik
Master's Thesis in Computer Science
von / by

André de Oliveira Melo

angefertigt unter der Leitung von / supervised by

Prof. Dr. Gerhard Weikum

betreut von / advised by

Dr. Martin Theobald

begutachtet von / reviewers

Dr. Max Mustermann

Prof. Dr. Gerhard Weikum

November / November 2012

Hilfsmittelerklärung

Hiermit versichere ich, die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt zu haben.

Non-plagiarism Statement

Hereby I confirm that this thesis is my own work and that I have documented all sources used.

Saarbrücken, den 15. November 2012,

(André de Oliveira Melo)

Einverständniserklärung

Ich bin damit einverstanden, dass meine (bestandene) Arbeit in beiden Versionen in die Bibliothek der Informatik aufgenommen und damit veröffentlichtwird.

Declaration of Consent

Herewith I agree that my thesis will be made available through the library of the Computer Science Department, Saarland University.

Saarbrücken, den 15. November 2012,

(André de Oliveira Melo)

To my father Cicero, my mother Marlene and my sister Carolina

- Andre

Abstract

With millions of articles in multiple languages, Wikipedia has become the de-facto source of reference on the Internet today. Each article on Wikipedia contains encyclopedic information about various topics (people, events, inventions, etc.) and implicitly represents an entity. Extracting the most important facts about such entity will help users to find desired information more quickly and effectively. However, this task is challenging due to the incomplete and noisy nature of Wikipedia articles. This calls for a mechanism to detect and summarize the most important information about an entity on Wikipedia.

This thesis proposes and implements CATE (**C**ontext-**A**ware **T**imeline for **E**ntity Exploration), a framework that utilizes Wikipedia to summarize and visualize the important aspects of entities in a timeline fashion. Such a system will help users to draw quickly an informative picture of an entity (e.g. life of a person, or evolution of a research topic, etc.). The novelty of CATE lies in seeing the entity in different contexts, synchronous with contemporaneous events. In addition, CATE puts the entity in a relationship with other entities, and thus offers a broader portrait about it. In order to efficiently query and visualize the events related to the entity, a number of techniques have been developed, combining information extraction and information retrieval with a novel ranking model. The thesis also discusses several experiments and evaluation results to show the effectiveness of the methods proposed.

Acknowledgements

Firstly, I would like to thank my advisor Dr. Martin Theobald, for his invaluable guidance. I feel deeply grateful for his technical assistance and motivational encouragement.

A special note of thanks to Prof. Gerhard Weikum for giving me the opportunity to pursue this thesis at Infomation and Database Systems department under his supervision. It was a very enriching and pleasant experience to write my Master Thesis here.

Contents

Chapter 1

Introduction

In the last years, the volume of semantic data available, in particular RDF, has dramatically increased. Initiatives like the W3C Semantic Web, which provides a common standard that allows data to be shared and reused across different applications, and the Linked Open Data, which provides linkages between different datasets that were not originally interconnected, have great contribution in such development. Moreover, advances in information extraction have also made strong contribution, by crawling multiple non-structured resources in the Web and extracting RDF facts.

Nevertheless, information extraction still has its limitations and many of sources might contain contradictory or uncertain information. Therefore, many of the extracted datasets suffer from incompleteness, noise and uncertainty.

In order to reduce such problems, one can apply to the knowledge base a set of inference rules that describes its domain. With that, it's possible to resolve contradictions or strengthen or weaken their confidence values. It's also possible to derive new facts that are originally not existent due to incompleteness. Such inference rules can be of two types:

1. *Hard Rules*: Consistency constraints which represent which might represent functional dependencies, functional or inverse-functional properties of predicates or Mutual exclusion. For example:

   marriedTo(x,y) :- marriedTo(y,x) grandChildOf(x,y) :- childOf(x,z)childOf(z,y)

2. *Soft Rules*: Weighted datalog rules that hold only for a limit amount of cases in real world. As they might also produce incorrect information, each rule itself must have a confidence value which should be applied to derived facts, for example married people live in the same place as their partner has confidence 0.9:

livesIn(x,y) :- marriedTo(x,z)livesIn(z,y) [0.9]

In such case, if we have an incomplete knowledge base, which lacks information about where *Michelle Obama*l lives, but we know that she's married to *Barack Obama* and he lives in *Washington, D.C.*, both with confidence 1, we could then apply this soft rule to derive the fact *livesIn(MichelleObama, Washington,D.C.)* with confidence 0.9.

data mining, rule mining, datalog rules

talk a bit about ilp

1.1 Motivation

Given the huge size of data, learning datalog rules can be already extremely costly.

1.2 Contributions

In order to make a web archive serve its many purposes, it is imperative to find efficient ways to control redundancy. This work makes the following contributions towards making this possible:

1. a framework that allows users to define redundancy based on (mutual) containment of document contents and meta data constraints,
2. an approach that determines for a user-defined notion of redundancy a small set of documents representative for the collection,
3. efficient algorithms to implement the approach on top of modern platforms for distributed data processing and management, specifically, Hadoop and HBase,
4. a end-to-end system prototype that can take a web archive document collection as input and produce a collection with redundancy removed.

1.3 Outline

The remainder of this thesis is structured as follows. In Chapter **??**, we provide technical background on MapReduce and BigTable. In Chapter **??**, we present a summary of

previous work in the areas of duplicate and near-duplicate detection, information retrieval on web archives, and MapReduce applications in graph processing. Following that, we state our problem and describe solutions in Chapter **??**. In Chapter **??**, we describe an implementation of our solution using the MapReduce framework. In Chapter **??**, we present our experimental results. We conclude this thesis and outline directions of future research in Chapter **??**.

List of Figures

List of Tables

Bibliography