WikiCat Browser: Exploring the Data from Wikipedia Categories’ Point of View

Project Description:

- Background:

  In loose terms, the goal in this project is to developed an interactive tool which utilizes the data annotations like extracted entities to enables users to explore the data based on their relation to the concepts of an external knowledge base, namely Wikipedia. This system helps users to dig into the data from their own perspective and explore which aspects of their interested concepts are related to a specific data.

  This project is a part of an already started project which aims to facilitate exploratory search in parliamentary data, ExPoSe, and the final goal of this part is to develop a system to project parliamentary debates on wikipedia categories. A preliminary version of the system has been developed and it needs improvements in some aspects including searchability and dynamicity.

  Here is some information on the developed version of the system:

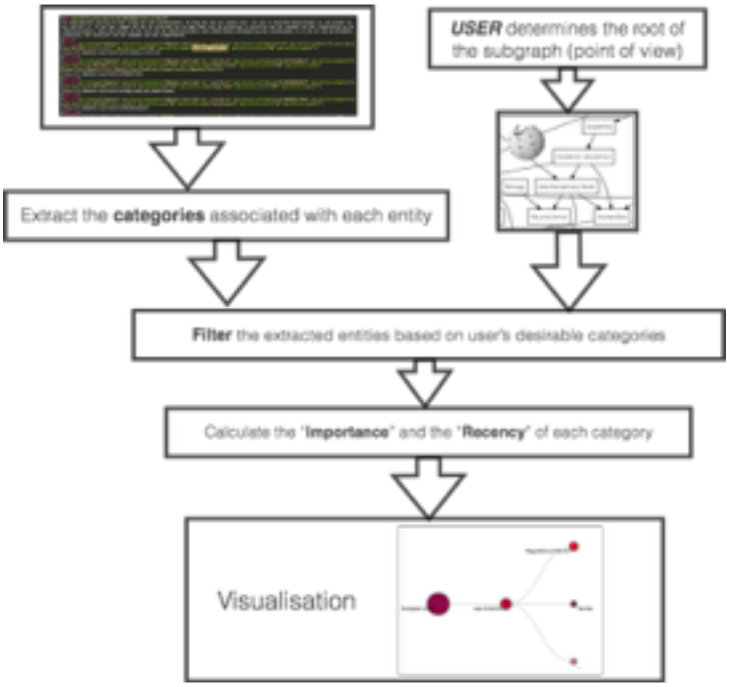
  ExPoSe WikiCat Browser is a system that enables users to browse and investigate the data from a particular point of view. The system makes use of extracted entities from the data in order to project the data on an arbitrary ontology. In fact, the structure of the selected ontology determines the point of view from which users want to see the data.

  As an example from the parliamentary domain, assume a user is interested in analysing the relation between national laws of a european country and EU legislation. One approach would be to investigate when EU legislation was discussed in the national parliament, and in the context of which (proposed) national laws. In other words, it is desirable to see how debates within the parliament of that country can be projected onto the topics related to the European Parliament in terms of both the subject matter and time.

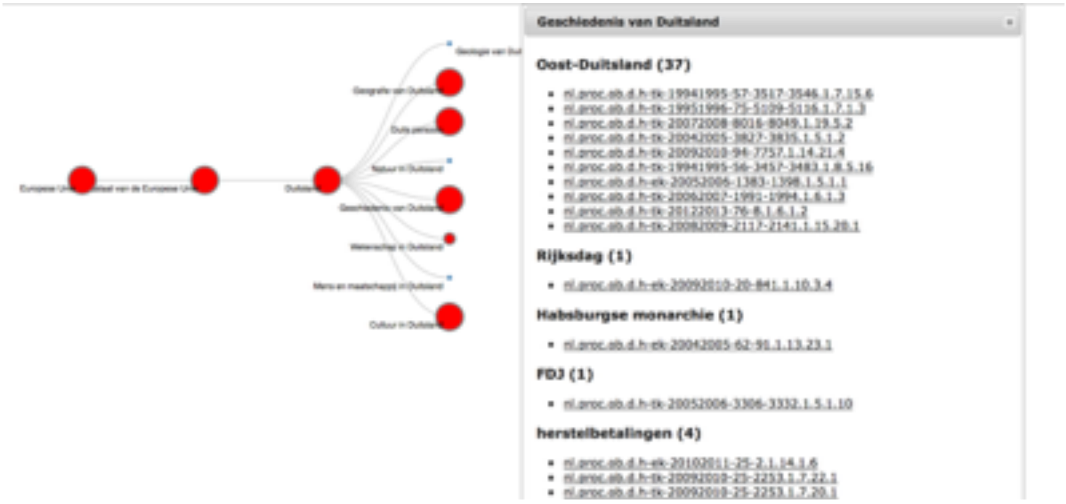
  In ExPoSe WikiCat Browser, we make use of the entity linker’s output to determine the notion of the topics of the data and we let the user select a sub-hierarchy from Wikipedia’s category hierarchy onto which mentioned entities from the text are projected. In the selected hierarchy, each node is a category which represents a topic and its descendant nodes are its sub-topics.

  In the parliamentary data example, the ExPoSe WikiCat Browser is provided with the extracted entities from national parliamentary debates and the “European Union” is selected as the category at the root of the hierarchy.

  Subsequently, the system extracts all the categories in the selected hierarchy and maps extracted entities to the Wikipedia categories. Afterward, it filters the entities considering whether they present some concepts related to the extracted categories or not. Then, having the debates and entities in them which are related to the categories in the hierarchy, the system calculates the **importance** and **recency** of each category (nodes in the hierarchy). The importance of each node demonstrates how much the topic of that category is addressed in the national parliamentary debates, based on frequency of entities related to this category. The recency of of each node shows how recently the topic of this category is discussed in the debates.


USER determines the root of the subgraph (point of view)
Extract the categories associated with each entity
Filter the extracted entities based on user's desirable categories
Calculate the "Importance" and the "Recency" of each category
Visualisation

A graphical user interface is provided from which the user is able to traverse the paths in the hierarchy and see which categories are more discussed in the national parliamentary debates (importance of nodes is shown by their size), which categories are recently discussed (recency of nodes is shown by their color), and which debates are related to which categories at different levels of abstraction. The links to all the debates related to the given category is provided grouped by the entities that belong to the category.



Geschiedenis van Duitsland

Oost-Duitsland (37)

- nl.proc.ob.d.h-tk-19941995-57-3517-3546.1.7.15.6
- nl.proc.ob.d.h-tk-19951996-75-5109-5116.1.7.1.3
- nl.proc.ob.d.h-tk-20072008-8016-8049.1.19.5.2
- nl.proc.ob.d.h-tk-20042005-3827-3835.1.5.1.2
- nl.proc.ob.d.h-tk-20092010-94-7757.1.14.21.4
- nl.proc.ob.d.h-tk-19941995-56-3457-3483.1.8.5.16
- nl.proc.ob.d.h-ek-20052006-1383-1398.1.5.1.1
- nl.proc.ob.d.h-tk-20062007-1991-1994.1.6.1.3
- nl.proc.ob.d.h-tk-20122013-76-8.1.6.1.2
- nl.proc.ob.d.h-tk-20082009-2117-2141.1.15.20.1

Rijksdag (1)

- nl.proc.ob.d.h-ek-20092010-20-841.1.10.3.4

Habsburgse monarchie (1)

- nl.proc.ob.d.h-ek-20042005-62-91.1.13.23.1

FDJ (1)

- nl.proc.ob.d.h-tk-20052006-3306-3332.1.5.1.10

herstelbetalingen (4)

- nl.proc.ob.d.h-ek-20102011-25-2.1.14.1.6
- nl.proc.ob.d.h-tk-20092010-25-2253.1.7.22.1
- nl.proc.ob.d.h-tk-20092010-25-2253.1.7.20.1

So, this system provides a hierarchical grouping of elements/documents in a large collection based on Wikipedia categories. In loose terms, it assigns categories to the documents employing the extracted entities from them and scores categories based on the time of documents assigned to them (recency) as well as the frequency of documents entities (importance). The system provides the user with a dynamic interface in which the categories, their recency/importance, and their relations are demonstrated and the user is able to browse and investigate the data from a very abstract level (categories) into a very detailed level (entities in documents).

The general idea of the system is to find the abstract representation of a data from a particular point of view, which is determined by the user, and empower the user to dive into the data from the abstract representation to the detailed information. This idea is applicable in other domains as well. For example, there are interesting discussion forums in social media that are growing and investigating and analysing this kind of data is getting more sophisticated. So, there would be desirable to have a system which not only represents the data from the point of view which is important for us, but also provides a dynamic way to browse the data in different levels of abstraction. Having these domains in mind, we are going to extend our system and evaluate it for other applications in other domains.

- Research Question(s):
  - How we can improve the searchability of the system?
  - How we can improve the dynamicity of the system?
  - How we can automatically evaluate the system?

- General Methodology:
  In the developed version of the system, exploring the data is from wikipedia categories to speeches in the debates and the user is able to choose a path in the hierarchy of categories to dive into data. However, sometimes it is not obvious how to traverse the hierarchy (how to choose the best path). In this situation, an idea is to get a query from the user and provide a ranking list of suggested paths. In other words, there is a search engine which instead of providing related documents, it gives paths which lead to related documents.
  Another challenge is how to evaluate the system. One approach is user study, however, there might be a way to automatically evaluate the system, or independent modules of the system.

Datasets:
- The target datasets are Dutch, Canadian, and British parliamentary proceedings which are already prepared and fully annotated in PoliticalMashup project.

Open Source code:
- The source code of the developed version of the system will be provided.

Outcome:
- The outcome of the project would be a new version of the system in which deficiencies are elaborated and the system will be presented as a demo in a conference (demo paper as output).

Literature:
- Documentation of the developed version.

Supervisor:
- Mostafa Dehghani (dehghani@uva.nl)