ON RECOMMENDING EVOLUTION MEASURES: A HUMAN-AWARE APPROACH

Kostas Stefanidis
Haridimos Kondylakis
Georgia Troullinou

Kostas Stefanidis

- University of Tampere, Finland

Haridimos Kondylakis

- ICS-FORTH, Greece

Georgia Troullinou

- ICS-FORTH, Greece

MOTIVATION

Over the past decade, numerous knowledge bases have been built to power largescale knowledge sharing, but also an entity-centric Web search

Knowledge bases offer comprehensive, machine-readable descriptions of a large variety of real-world entities published on the Web as Linked Data

Dynamicity is an indispensable part of the Linked Data

- Datasets are constantly evolving due to new experimental evidence or observations, or correction of erroneous conceptualizations

HOW TO UNDERSTAND THE KBS EVOLUTION

Study KBs deltas between different versions

Important task for several tasks:

- Synchronization of autonomously developed dataset versions
- Visualization of the evolution history of a dataset
- Accessing previous versions of a dataset to support historical or cross-snapshot queries
- Integrate and synchronize interconnected Linked Data

HOW TO UNDERSTAND THE KBS EVOLUTION

Capture important characteristics for quantifying the intensity of the changes that a knowledge base underwent

One can assume as important the actual number of changes that the classes or properties of a KB underwent during the evolution process

Number of class or property changes

Low-level deltas for counting the triples added and deleted during evolution

○ When interested in changes related to a particular class or property, take into account the triples referring to that class or property

HOW TO UNDERSTAND THE KBS EVOLUTION

Number of class or property changes in neighborhoods

E.g., when studying the evolution of a class n, we may be interested in the classes around n

- Determine whether the topology of the KB changed in a particular area

The neighborhood of a class n consists of the classes related to n via a subsumption relationship, or connected with n via a property

The number of changes in the neighborhood of n is equal to the sum of the changes of the nodes in the neighborhood

HOW TO UNDERSTAND THE KBS EVOLUTION

Structural measures

Bridging Centrality: identify the topological locality of a node in a graph

- A node with high *Bridging Centrality* connects densely connected components in a graph

Betweenness: count the number of paths between all nodes pairs that pass through a particular node

A shift in a node's Bridging Centrality or Betweenness from V1 to V2 could capture how changes on a dataset affected the topology around this specific node

HOW TO UNDERSTAND THE KBS EVOLUTION

Semantic measures

Centrality: quantify how central is a class n within a version

- *Relative Cardinality* of a property e(n, n_i): # of instance connections between n and n_i divided by the total # of instance connections the two classes have
- *Centrality*: sum of the (weighted) relative cardinalities of the properties

Relevance extends centrality to consider neighborhoods

The relevance of a class is affected by the centrality of the class itself, as well as by the centrality of its neighboring classes

OUR GOAL

Many different views of evolution that we could consider!

But, without requiring much work, how to help users:

- Get a supervisory overview of the changes
- Observe changes trends
- Identify the most changed parts of a knowledge base

Recommendations

- Suggest evolution measures, or their mix, that allow quantifying the changes that particular parts of a knowledge base underwent, and cover complementary viewpoints

OUR GOAL

Put humans in the core!

- Humans generate and consume huge amounts of data, and are interested to be notified about how data evolve
  - Social networks, sensors on the roads, online transactions

Exploit evolution measures to suggest different ways to understand and realize how data evolve and which are the main changes

DATA FOR & ABOUT PEOPLE

Power:

- Enormous data sets, enormous computational power, massively parallel processing

Opportunity:

- Improve people's lives - recommendations
- Accelerate scientific discovery - medicine
- Boost innovation - autonomous cars
- Transform society - open government
- Optimize business - advertisement targeting

DATA RESPONSIBILITY

The problem is not only in the technology, but also in how its used

Because of its tremendous power, massive data analysis must be used responsibly

- Relatedness
- Transparency
- Diversity
- Fairness
- Anonymity

RELATEDNESS

Exploring the contents of a KB to study how it evolves is a complex process that may return a huge volume of data

- Consider the abundance of available information

Users would like to retrieve only a small piece of the evolved data

- The most relevant to their interests and needs

Not enough work has been done towards associating the relatedness of the evolution of specific parts of a KB with humans

TRANSPARENCY

Help humans to know what is being recorded for them and the evolution process, and how the recorded information is being used

Provenance information is important for achieving transparency

- Who created this data item and when, by whom was the data item modified and when, what was the processes used to create the data item

Consider how to support the automation of repetitive tasks, and systematically capture provenance information

- Care about the truth of the provenance data

DIVERSITY

A big number of studies motivate the benefits diversity provides

The challenge: introduce algorithms resulting in sets of evolution measures that as a whole exhibit a desired property

- Not just assign individually interest scores to measures

The produced set of measures should cover all the different needs of the human in question and not focus on a particular aspect of evolution

FAIRNESS

Lack of bias: bias can come from data processing methods that reflect the preferences of the scientists designing them – Support uncommon information needs

Group notion of fairness: E.g., recommend evolution measures to the curators' team of a KB

- Locate suggestions that include measures fair to the members of the group

In actual life, we should be able to recommend measures that are both strongly related and fair to the majority of the group members

- Have insights into the properties of the produced recommendations to help making the algorithmic process non-discriminative

ANONYMITY

To observe data evolution: Find patterns that usually happen, and perform some aggregations on them

- This is a method for achieving anonymity as well

Somehow, studying data evolution, suffices privacy issues

- Often, even if data is aggregated, it is possible to re-identify sensitive data or significant parts of it

Access to personal and private data is essential, meaning that strict rules prohibiting reach such data should apply

OUR GOAL

We target at introducing a recommenders-like way to assess the evolution intensity of KBs

This is intended as an aid for the humans, allowing them to quickly understand how data changes and get an overview of the important changes under different perspectives

THANK YOU!

Questions?