ECIS 2017 - Similarity Workshop

June 6th, 2017 - Guimarães, Portugal

Overview

The increasing availability of digital trace data and recent advances on both methods and tools to analyze data in unprecedented scale open up new venues for information systems research. These trends permit the implementation of novel analysis methods and create new opportunities for theorization.

Yet, the sheer volume of data means that the traditional research methods, whether qualitative or quantitative, may not be able to fully benefit from the depth and the breadth of the data available. That is unless the data is made accessible through intermediary transformation steps.

Traditionally this accessibility has been achieved by limiting the scale of the data through sampling strategies.

But advances in computational methods have made other strategies possible that may provide additional or holistic insights. Among others, these methods include dimensionality reduction methods used to "summarize" the data by focusing on its salient features and the methods capable of detecting patterns and similarities at a large scale. The combination of these computationally-powered tools with traditional research methods create new possibilities for both inductive and deductive research.

This workshop focuses on similarity detection methods. An interactive workshop will be held where the participants will have the chance to use similarity detection methods on pre-collected data. The workshop will include presentations and interactive sessions focusing on the following methods:

- Locality-sensitive hashing for text similarity detection (Mahmood Shafeie Zargar).
- Spherical harmonic representation of 3D shapes for product design comparison (Harris Kyriakou).
- Semantic transformations with ontology extraction using Wikipedia (Yegin Genc).

Methods

Locality-sensitive hashing

Locality-sensitive hashing techniques are used to compare distinctly different items and quantifying their degree of similarity in form of a similarity score or percentage. Unlike cryptographic hashes that are designed to identify exact duplicates of items, locality-sensitive hashes are capable of identify partial matches between items. Locality-sensitive hashes are in widespread use for malware and spam detection, plagiarism detection, audio and video fingerprinting, as well as copyright enforcement.

Among the different families of locality-sensitive hashes, in this workshop we focus on fuzzy hashing, or context triggered piecewise hashing (CPTH), originally used for detecting emails with similar content.[1,2]

Given that fuzzy hashes can detect similarities between texts, they are an ideal tool for tracking down the diffusion and evolution of discourse in social media.

[1] Tridgell, A. (1999). *Efficient algorithms for sorting and synchronization* (PhD). The Australian National University, Canberra, Australia.

[2] Kornblum, J. (2006). Identifying almost identical files using context triggered piecewise hashing. *Digital Investigation*, *3*(Supplement), 91–97.

Rotation & Scale Invariant Spherical Harmonic Representation

A rotation invariant spherical harmonic representation method to measure the similarity between product designs will be presented. The method is based upon a variation of a computer graphics method for calculating the shape distance between 3D shapes.[3] The algorithm represents each 3D design based on spherical harmonics, in order to obtain rotation and scale invariant characterizations that can be used to calculate distances that represent changes in shape rather than changes in perspective.

One way to conceptually understand the technique is to imagine hollow 3D objects, and consider filling these objects with some number of tennis balls, ping pong balls, and ball bearings. Objects that are similar will need a similar proportion of balls of different sizes to fill them up.

The workshop will showcase the use of a 3D shape comparison method as well as touch upon the similarity measures based upon the semantic description of designs. Such methods can provide insights about the similarities of products when data from other methods (e.g. text, network structure) are not available, or complement our insights from text and network based methods.

[3] Kazhdan, M., Funkhouser, T., & Rusinkiewicz, S. (2003). Rotation invariant spherical harmonic representation of 3 d shape descriptors. In *Symposium on geometry processing*, *6*, 156–164.

Ontology Extraction

In order to effectively measure the semantic similarity of documents, we need to take into account the context. Ontologies can be used to model concepts and their relationships. In this sense, ontologies can be used to represent relevant aspects of context. However it is not a trivial task to detect whether a word, or a part-of-text, in a document is a concept.

The Ontology Extract method aims to detect the concepts covered in a document by using Wikipedia as the external knowledge domain. Wikipedia's human readable labels to the concepts help semantic analysis in the Information Retrieval Tasks.[4]

Wikipedia based ontologies can be specially useful when analyzing short and elliptic text. For example, when classifying tweets into their topics, ontologies extracted from Wikipedia can be more helpful in identifying the text similarity than other statistical methods that rely on word frequencies.

[4] Gabrilovich, Evgeniy, and Shaul Markovitch. (2009) Wikipedia-based semantic interpretation for natural language processing. *Journal of Artificial Intelligence Research*, *34*, 443-498.

Program

Workshop Date: June 6th 2017 (Half-day Workshop)

08:30-09:00	Registration & Coffee
09:00-09:15	Introduction & Overview of the Workshop
09:15-09:45	Showcase 1: Similarity hashing on programming code
09:45-10:15	Showcase 2: Shape comparison on representations of 3D printed product designs
10:15-10:45	Showcase 3: Ontology extraction using Wikipedia
10:45-11:10	Break
11:10-	

12:10	Hands-on Break-out Sessions
12:10-12:30	Q&A, Concluding Remarks, Feedback and Informal Discussion

Participation

If you would like to participate in the workshop, please submit a participation request. You will receive a confirmation within a few hours.

Requirements

There is no requirement for workshop participation, although some programming background (Python) is necessary for active learning during the hands-on session. The participants have to bring their own personal computers for the hands-on session. It is also the participants' responsibility to install the required set of free software tools and libraries mentioned under the resources section prior to the workshop.

Organizers

Mahmood Shafeie Zargar (Coordinator)
Assistant Professor
Vrije Universiteit Amsterdam
Netherlands
Website | Email

Harris Kyriakou
Assistant Professor
IESE Business School
Spain
Website | Email

Yegin Genc
Assistant Professor
Pace University
USA
Website | Email