Universiteit Utrecht

[Faculty of **Science**
Information and Computing Sciences]

Hybrid approaches to software clustering: from the human in the loop to combining multiple sources of information

Jurriaan Hage >>= Amir M. Saeidi

Utrecht University, Mendix

April 12, 2016

1. Machine Learning for Clustering

Applications of ML in Software Engineering §1

- Many software development and maintenance tasks can be formulated as learning problems
- Applications of ML in software engineering include
  - prediction and estimation
  - property and model discovery
  - transformation
  - generation and synthesis
  - reuse
  - requirement acquisition
  - management of development knowledge
- We adapted ML algorithms to combine different sources of information and/or allow for human intervention.
- Today we look at clustering of software systems

(Semi-)supervised Clustering

- Software clustering is the task of organizing software units into modules
  - The notion of "module" is flexible
- Applications of clustering in the context of source code analysis:
  1. Obtain a high-level overview of a software system consisting of software clusters
  2. Determine how much the architecture of the system has drifted
  3. High-level overview is necessary to enable further, more detailed tasks
  4. Often a first step in a software comprehension setting

Types of Clustering

- There are three major approaches to clustering:
  1. structural - file organization or call graph
  2. lexical (or semantic) - vocabulary used in the source code
  3. evolutionary - information obtained from the revision history (eg. changes simultaneously submitted to the version archive by the same developer, or developer team or file ownership information)
- The structural and lexical-based clustering have shown to outperform the evolutionary approaches.

Structure-based Clustering

- A fundamental assumption underlying this approach is that well-designed software systems are organized into cohesive clusters (high-cohesion) that are loosely interconnected (low-coupling).
- **Objective**: minimize inter-connectivity (i.e., connections between the components of two distinct clusters) while maximizing intra-connectivity
- Modularization Quality (MQ) is defined as a trade-off between inter-connectivity and intra-connectivity
- Local search algorithms such as hill climbing or genetic algorithms to find 'good' solutions

Semantic Clustering

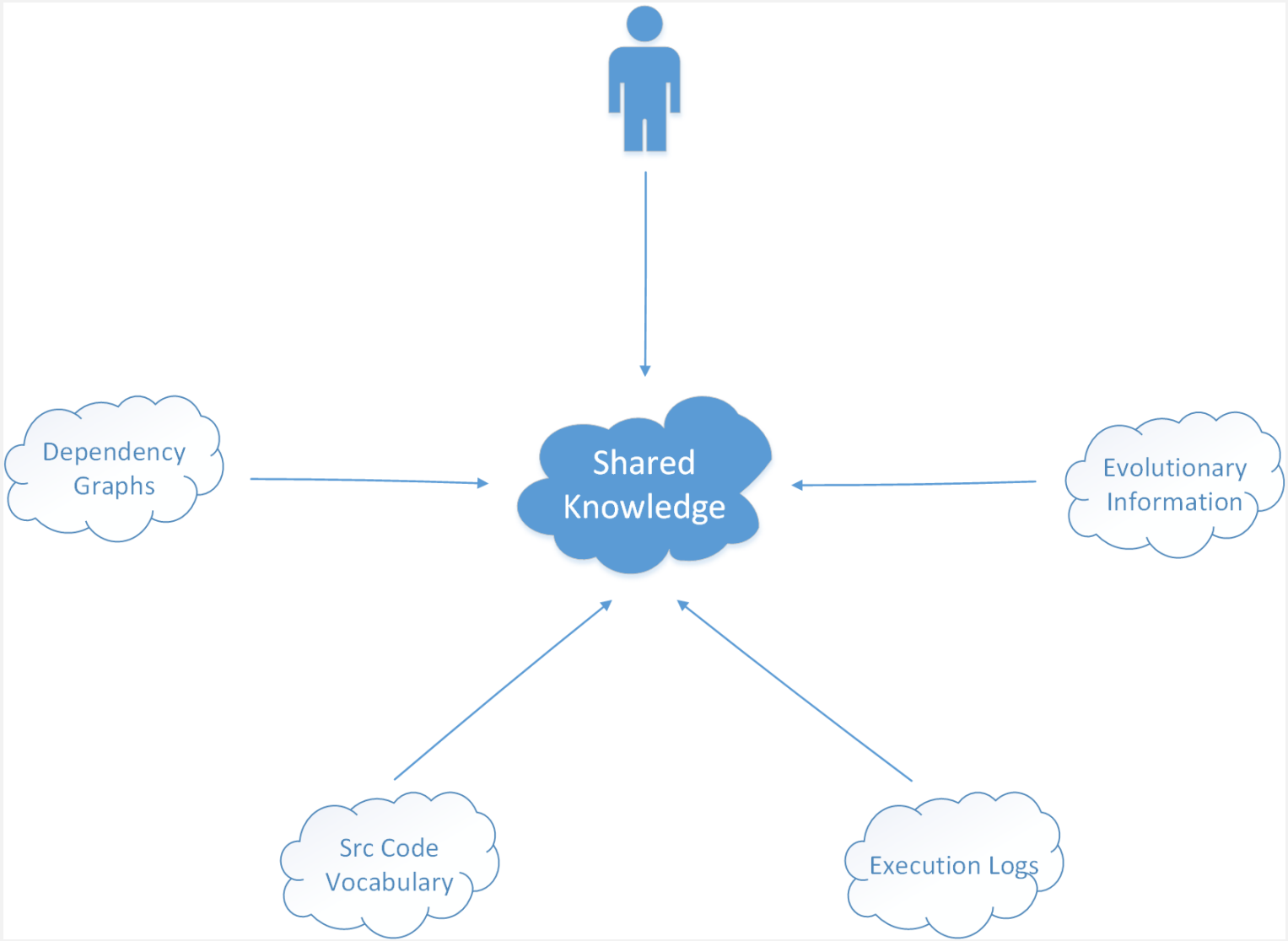
- The informal semantics of the source code is concealed within comments, identifier names and/or literals
- Semantic clustering uses information retrieval techniques to:
  1. partition the system into clusters based on the use of similar vocabulary
  2. derive a set of common linguistic topics within each cluster
- Semantic clustering has been shown in various studies to capture important domain and application concepts of a software system

2. Hybrid Approaches to Clustering

Multi-View Data

- Different views of objects, each have specific statistical and geometrical properties
- Example: a document available in Spanish, Finnish and Dutch, each language is a "view"
- Combine views to construct a new translation to resolve issues of synonymy (multiple words with same meaning) and polysemy (same word with multiple meanings).
- E.g., a word in Finnish may have only one meaning, but its translation into Dutch may have more. For other words, the situation may be reversed.

Combining different sources of information



Multi-View Clustering

- Principles used for successful clustering of multi-view data:

  Consensus : maximize the agreement on multiple distinct views.

  Complementary : each view of the data may contain some knowledge that other views do not have; therefore, multiple views can be employed to comprehensively and accurately describe the data.

Results

Empirical evaluation of 10 open source Java projects [Saeidi et al.]

- Structural dependencies and semantic similarity are more reliable sources of information than evolutionary information.
- As systems grow in size, the (single-view) heuristics used for remodularization are not the driving factors

Results - cont'd

- Multi-view clustering yields decompositions that are at least as authoritative as the least informative view
  - Nothing is lost by using multi-view approaches.
- In most cases, multi-view clustering outperforms single-view setting
- Multi-objective encoding of multi-view clustering can be used to obtain a pool of solutions (Alternative Cluster Analysis)
  - An expert may choose the one which aligns with his/her expectation of what the modularity of the system should be like.

3. Interactive Topic Modeling for Software Understanding

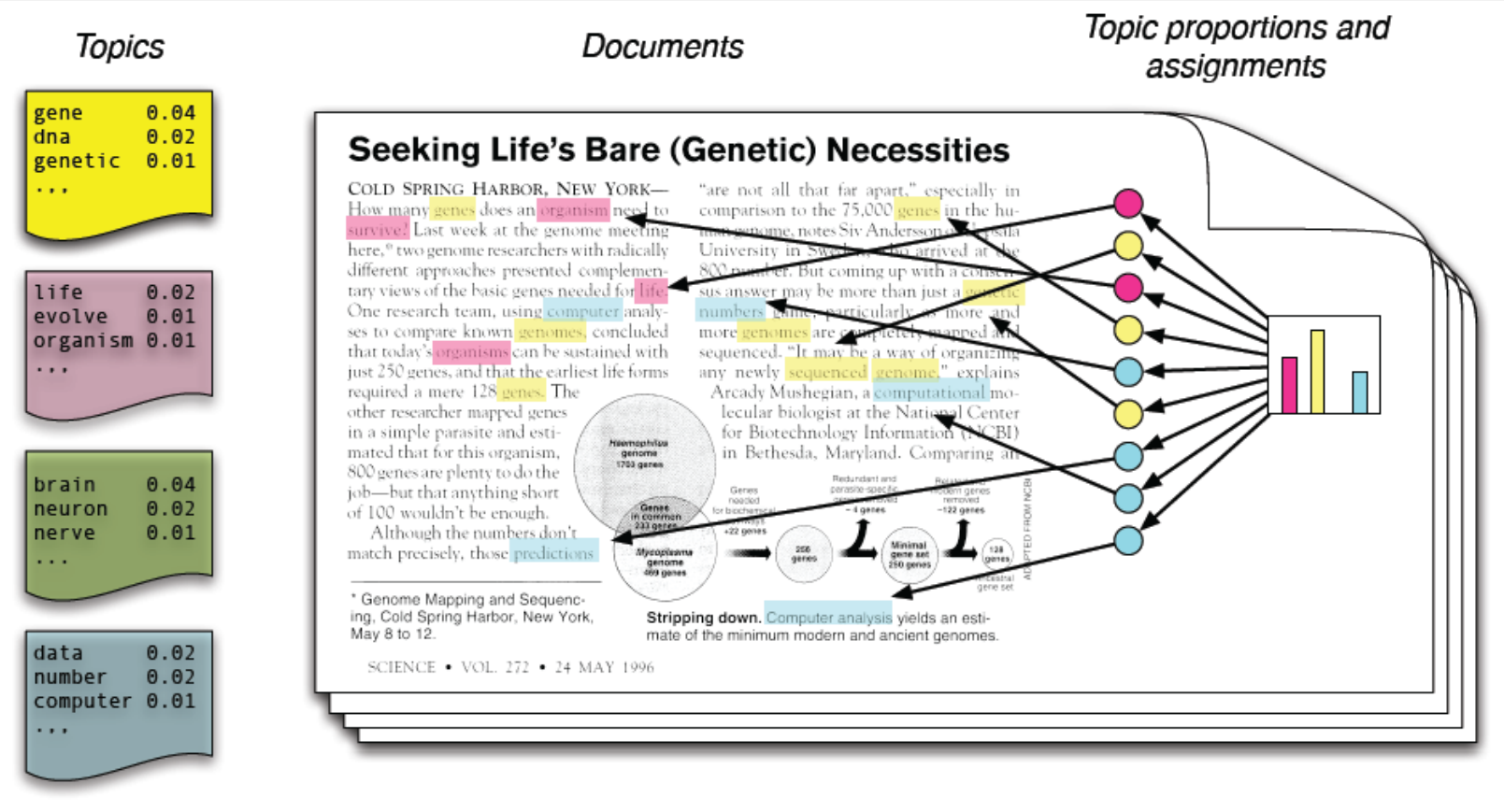
Topic Modeling

- A topic model is a type of statistical model for discovering the abstract "topics" that occur in a collection of documents
- Used for analysis and discovery of meta-data of the documents in a wide range of domains: web pages, images, sounds and video, genetics, social networks
- Goals:
  - discover the themes that exist within the documents
  - how those themes are connected to each other
  - how they change over time
- Topic models' picture of the outside world is based on "bag-of-words"

Bag-of-Words

- Segmentation of all composite terms, for instance delimited by '-'
- Elimination of common terms that occur in a natural language such as "de", "het" and "een" in the case of Dutch
- Removal of programming language-specific reserved words such as PROCEDURE, DIVISION for Cobol
- Normalization of all lexical term by a natural language-specific stemmer to emit a common radix
- A weighing mechanism is applied to punish words that appear in many documents
- The resulting set of terms with its number of occurrences is stored as a "bag-of-words"

How does it work?



[1]

- Assume some k number of "topics" exists for the whole document
- Inference: assign words to each topic with a probability

[1] David Blei, Introduction to Probabilistic Topic Models

Source Code Naming

- Success of topic modeling to derive the vocabulary heavily depends on source code naming, and presence of comments
- Can deal with some level of noise/arbitrarity
- Coded format such as the abbreviated form does not affect performance

Interactive Topic Modeling

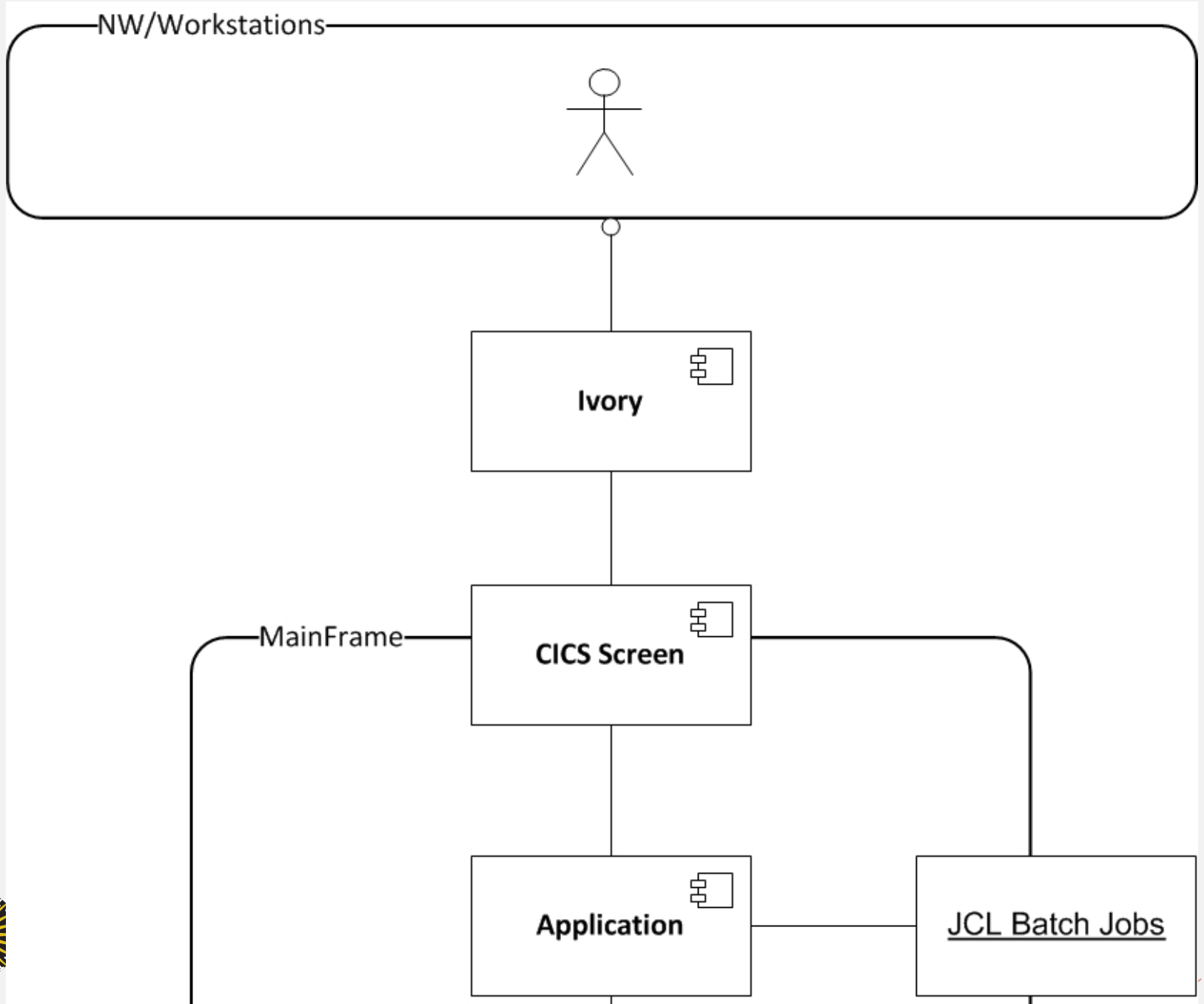
- Incorporates human knowledge to improve the topic models
- Allows users (architect, developers) to iteratively refine the topics discovered by topic modeling techniques
- Soft constraints: Must-link (words should share a topic) and Cannot-link (shouldn't)

Identifying Topics

- Associate domain and application concepts with terms, by consulting the system architect as well as developers
- Not all correspondences need to be established: a few might suffice to uncover the meaning of a given topic (= set of words grouped together)

4. ITMViz

ITMViz Architecture



Public Acess

- LDAVis: `https://github.com/cpsievert/LDAvis`
- Interactive LDA (ITM) package: `https://github.com/amirms/ITM`
- ITMViz application source code: `https://github.com/amirms/ITMViz`
- ITMViz running example: `https://gelato.shinyapps.io/ITMViz`

5. Questions