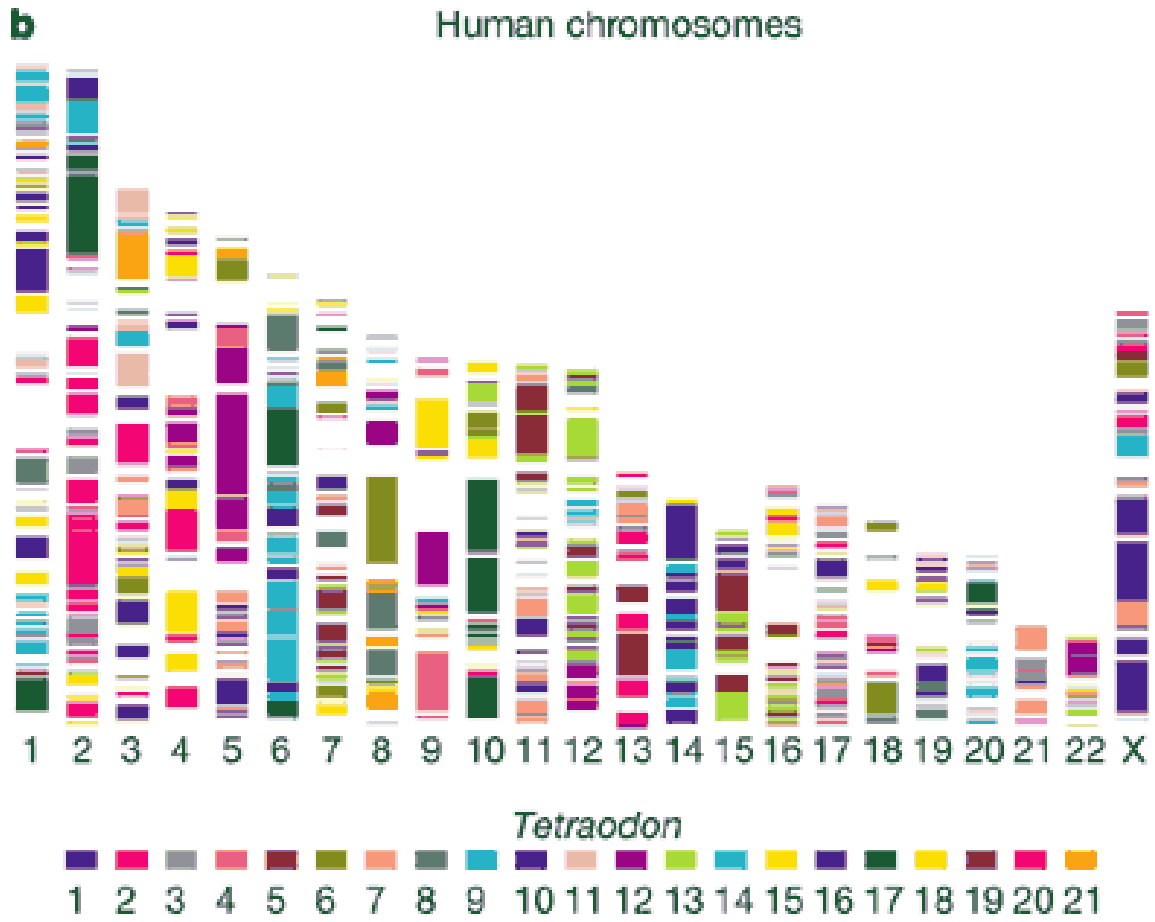
Similarity Search

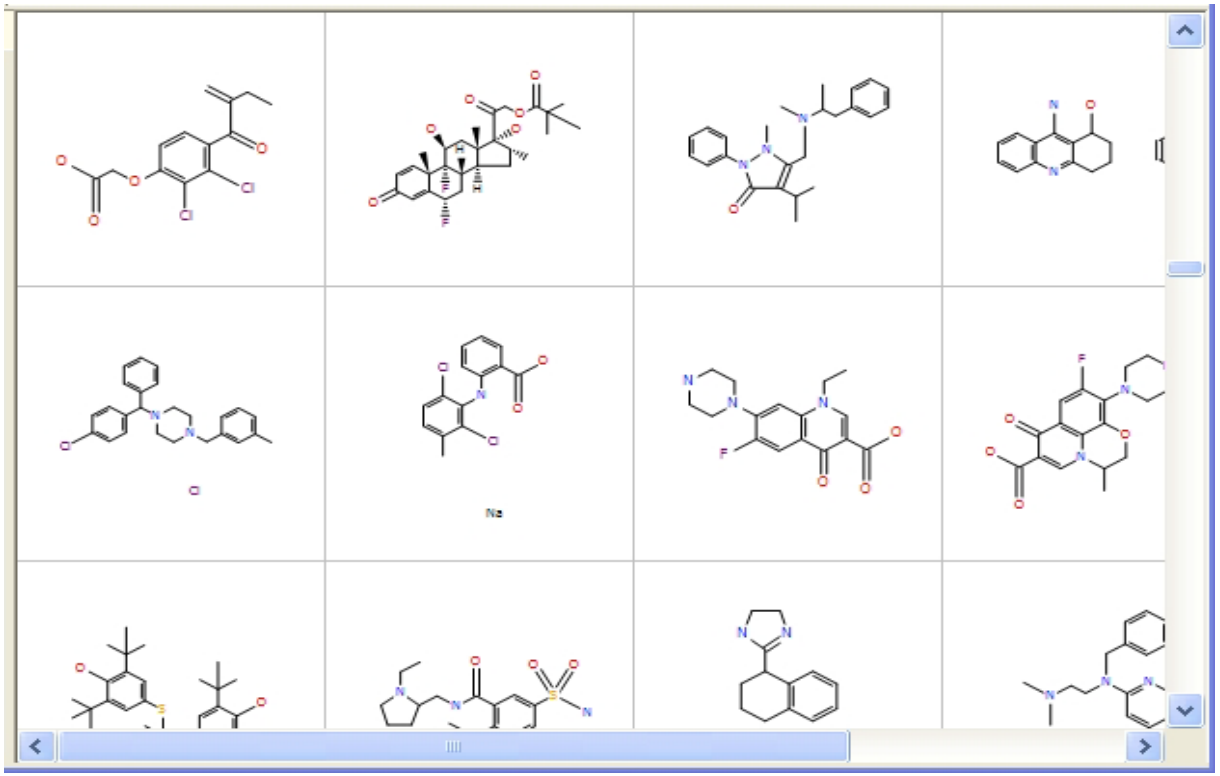
Ulf Leser

Similarity Search - Images

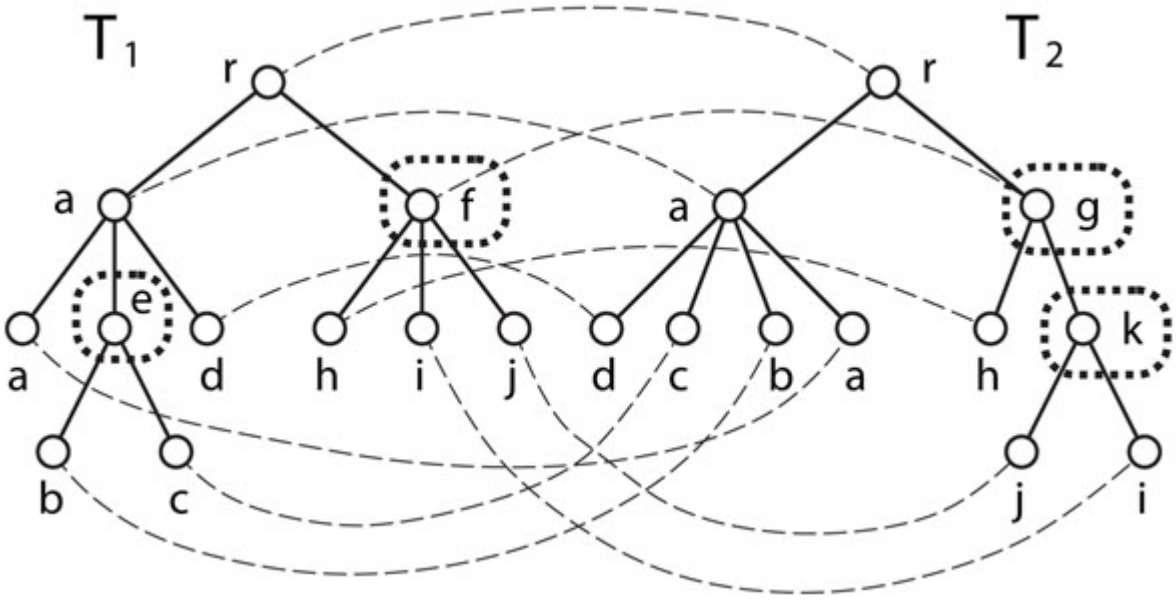
Similarity Search - Strings



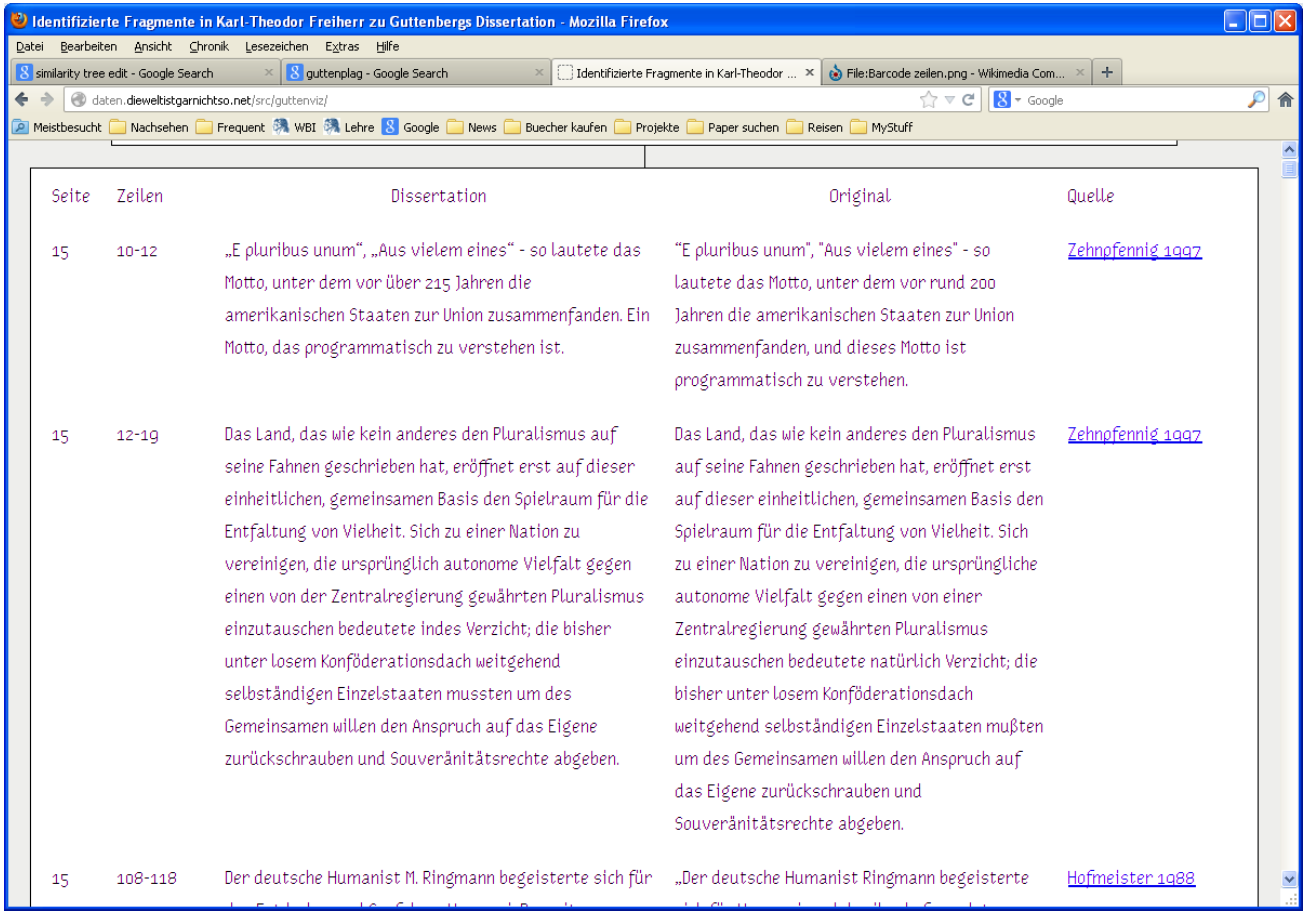
Similarity Search - Graphs



Similarity Search - Trees



Similarity Search - Documents



Seite	Zeilen	Dissertation	Original	Quelle
15	10-12	„E pluribus unum", „Aus vielem eines" - so lautete das Motto, unter dem vor über 215 Jahren die amerikanischen Staaten zur Union zusammenfanden. Ein Motto, das programmatisch zu verstehen ist.	"E pluribus unum", "Aus vielem eines" - so lautete das Motto, unter dem vor rund 200 Jahren die amerikanischen Staaten zur Union zusammenfanden, und dieses Motto ist programmatisch zu verstehen.	Zehnpfennig 1997
15	12-19	Das Land, das wie kein anderes den Pluralismus auf seine Fahnen geschrieben hat, eröffnet erst auf dieser einheitlichen, gemeinsamen Basis den Spielraum für die Entfaltung von Vielheit. Sich zu einer Nation zu vereinigen, die ursprünglich autonome Vielfalt gegen einen von der Zentralregierung gewährten Pluralismus einzutauschen bedeutete indes Verzicht; die bisher unter losem Konföderationsdach weitgehend selbständigen Einzelstaaten mussten um des Gemeinsamen willen den Anspruch auf das Eigene zurückschrauben und Souveränitätsrechte abgeben.	Das Land, das wie kein anderes den Pluralismus auf seine Fahnen geschrieben hat, eröffnet erst auf dieser einheitlichen, gemeinsamen Basis den Spielraum für die Entfaltung von Vielheit. Sich zu einer Nation zu vereinigen, die ursprüngliche autonome Vielfalt gegen einen von einer Zentralregierung gewährten Pluralismus einzutauschen bedeutete natürlich Verzicht; die bisher unter losem Konföderationsdach weitgehend selbständigen Einzelstaaten mußten um des Gemeinsamen willen den Anspruch auf das Eigene zurückschrauben und Souveränitätsrechte abgeben.	Zehnpfennig 1997
15	108-118	Der deutsche Humanist M. Ringmann begeisterte sich für	„Der deutsche Humanist Ringmann begeisterte	Hofmeister 1988

Who should be here

- Mono- / Kombibachelor Informatik or Infomit
- Ability to read English papers
- Good knowledge in databases and algorithms
  - Optimization, index structures, B-Trees, hashing, …
- Willingness to independent work
  - Search suitable papers covering a topic, prepare presentations, write seminar thesis

How it will work

- Today: Presentation and choice of topics
  - If desired, we will group teams of 2 students
- 1.12.2019: Send an outline of your topic (next slide)
- ~20.12.19: Present your topic in 5min flash-presentation
- ~25.1.20: Meet your advisor to discuss slides
- ~15.2.20: Present your topic (30min) at the Blockseminar
- 1.4.2020: Write seminar thesis (10-15 pages)

The outline

- Topics are rather abstract
- Find a set of suitable papers covering the topic
  - Focus is allowed and welcome
- Extract the most important information
- Structure into an outline of your thesis
  - Chapters, Sections, 1-2 sentences per section to describe the content, 2-4 figures
- Write an abstract
  - Roughly 20 lines – what is the topic, what will the thesis describe?
- Send me abstract + outline + references

The 5-min flash talk

- Focus on marketing – drag students into your topic
  - What is the topic?
  - Why is it challenging?
  - Why is it cool?
  - What are important applications?
  - What will your talk be about?
- At most 5 slides
- Focus on figures & examples; omit all details or algorithms

Presentation

- 30min presentation
- German or English
- Explain topic, methods, experimental results
- Compare different approaches (if enough time)
- Aim: Your audience should understand what you say
- No need to cover the topic entirely – select best topics

Teams

- If a topic is addressed by a team of two students, I expect
  - Read more papers
  - Have more topics in your outline and thesis
  - Presentations times remain the same – choose wisely

ToC

- Introduction
- Topics
- Assignment
- Hints on presenting your topic and writing your thesis

Topic	Assigned to
String (DNA) similarity	
Sentence similarity	
Text similarity	
Web page similarity	
Tree Similarity Search	
Graph similarity and graph alignment	
Time series similarity	
Music similarity	
Image similarity	
Workflow similarity	

Flavors

- How are the objects under study defined?
- What similarity / distance functions exist?
- What properties do they have?
- How can we compute the distance between two objects?
- Do we compare entire objects or just parts (containment)?
- How can we find the closest objects given a query object?
- How can we perform a similarity join over such objects?
- How can indexes support such operations?
- Do we compute exact solutions or approximations?
- Any formal bounds or "just" heuristics?
- If there are different methods – which is best?

Some properties 1

Topic	Assigned to
String (DNA) similarity	
Sentence similarity	
Text similarity	
Web page similarity	
Tree Similarity Search	
Graph similarity and graph alignment	
Time series similarity	
Music similarity	
Image similarity	
Workflow similarity	

- DNA sequences
  - Have an evolutionary explanation
  - Modern applications need to search petabytes of data
- Sentence similarity
  - Consist of words with semantics
  - But sentences a short – little context
- Text similarity
  - Entire texts or just parts (plagiarism)?
- Web page similarity
  - Many web pages!
  - Menus, ads, … should be ignored

Some properties 2

Topic	Assigned to
String (DNA) similarity	
Sentence similarity	
Text similarity	
Web page similarity	
Tree Similarity Search	
Graph similarity and graph alignment	
Time series similarity	
Music similarity	
Image similarity	
Workflow similarity	

- Trees
  - Can be ordered or not
  - Already tree edit distance is a challenge
- Graph similarity
  - Very high complexity (sub/graph isomorphism)
- Time series similarity
  - Time series of real values
  - Similarity is not easy to define
- Music similarity
  - Has structure and multiple channels

Some properties 3

Topic	Assigned to
String (DNA) similarity	
Sentence similarity	
Text similarity	
Web page similarity	
Tree Similarity Search	
Graph similarity and graph alignment	
Time series similarity	
Music similarity	
Image similarity	
Workflow similarity	

- Image similarity
  - Very difficult to define
  - Millions of pixels
- Workflow similarity
  - Nodes (programs) and edges (dependencies)

Topic	Assigned to
String (DNA) similarity	Eberlein
Sentence similarity	Arndt, Patzak
Text similarity	Day
Web page similarity	Sergelen, Othegraven
Tree Similarity Search	Salek, Riese
Graph similarity and graph alignment	Beiker, Weber
Time series similarity	
Music similarity	Kröger, Bektas
Image similarity	Zierle, Lahn
Workflow similarity	
Video similarity	Becker

ToC

- Introduction
- Topics
- Assignment
- Hints on presenting your topic and writing your thesis

Allgemeine Hinweise

- Dozenten sind ansprechbar!
  - Vorbesprechung des Themas
  - Folien durchgehen
  - Abgrenzung der Ausarbeitung
- Diskussion erwünscht
  - Keine Angst vor Fragen: Fragen sind keine Kritik
  - Eine Frage nicht beantworten können ist in Ordnung
- Tiefe, nicht Breite
  - Lieber das Thema einengen und dafür Details erklären
- Bezug nehmen
  - Vergleich zu anderen Arbeiten (im Seminar)

Allgemeine Hinweise

- Werten und bewerten
  - Keine Angst vor nicht ganz zutreffenden Aussagen – solange gute Gründe vorhanden sind
  - Begründen und argumentieren
  - Kritikloses Abschreiben ist fehl am Platz
- Literaturrecherche ist notwendig
  - Die ausgegebenen Arbeiten sind Anker
  - Weiterführende Arbeiten müssen herangezogen werden
  - Auch Grundlagen nachlesen

- Wir schicken eine Liste zum Abhaken rum

Wie halte ich einen Seminarvortrag

1. Wenn man nun so einen Seminarvortrag halten muss, dann empfiehlt es sich, möglichst lange Sätze auf die Folien zu schreiben, damit die Zuhörer nach dem Vortrag aus den Folienkopien noch wissen, was man eigentlich gesagt hat.

2. Während so einem Vortrag schaut sowieso jeder zum Projektor, also kann man das selbst ruhig auch tun - damit kontrolliert man gleichzeitig auch, ob der Beamer wirklich alles projeziert, was auf dem Laptop zu sehen ist. Ausserdem kann man so den Strom für das Laptop-Display sparen.

3. Übersichtsfolien am Anfang sind langweilig, enthalten keinen Inhalt und nehmen den Zuhörern die ganze Spannung. Schliesslich gibt‘s im Kino am Anfang auch keine Inhaltsangabe.

4. Powerpoint kann viele lustige Effekte, hat tolle Designs und Animationen. Die sollte man zur Auflockerung des Vortrags unbedingt alle benutzen, um zu zeigen, wie gut man das Tool im Griff hat.

5. Nicht zu wenig auf die Folien schreiben. Man weiß ja nie, ob man sie nicht doch ausdrucken muss, und man kann so wertvolle Zeit sparen, wenn man nicht weiterschalten muss.

6. Man sollte versuchen, möglichst lange zu reden. Die Zeitvorgaben sind nur für die Leute, die nicht genug wissen - eigentlich will der Prüfer sehen, dass man sich auch darüber hinaus mit dem Thema beschäftigt hat.

Bloß keine Hervorhebungen im Text – sonst müssen die Zuhörer ja gar nicht mehr aufpassen!

Hinweise zum Vortrag

- ~30 Minuten inkl Diskussion
- Klare Gliederung
- Ab und an Hinweise geben, wo man sich befindet
- Bilder und Grafiken; Beispiele
- Font: mind. 16pt
- Eher Stichwörter als lange Sätze
- Vorträge können auch unterhaltend sein
  - Gimmicks, Rhythmuswechsel, Einbeziehen der Zuhörer, etc.
- Adressat sind alle Teilnehmer, nicht nur die Betreuer
- Technik: Laptop? Powerpoint?

Hinweise zur Ausarbeitung

- Eine gedruckte Version abgeben
  - Selbstständigkeitserklärung unterschreiben
- Eine elektronische Version schicken
- Referenzen: Alle verwendeten und nur die
  - Im Text referenzieren, Liste am Schluss
- Korrekt zitieren
  - Vorsicht vor Übernahme von kompletten Textpassagen; wenn, dann deutlich kennzeichnen
  - Aussagen mit Evidenz oder Verweis auf Literatur versehen
- Verwendung von gefundenen Arbeiten im Web
  - Möglich, aber VORSICHT
  - Eventuell Themenschwerpunkt verschieben – Betreuer fragen

Format

- Benutzung unserer Latex-Vorlage
- Nur eine Schriftart, wenig und konsistente Wechsel in Schriftgröße und –stärke
- Inhaltsverzeichnis
- Bilder: Nummerieren und darauf verweisen
- Referenzen:
  - [1] Yan, X., Yu, P. S. and Han, J. (2004). "Graph Indexing: A Frequent Structure-Based Approach". SIGMOD, Paris, France.
  - [YYH04] Yan, X., Yu, P. S. and Han, J. (2004). "Graph Indexing: A Frequent Structure-Based Approach". SIGMOD, Paris, France.
- Darf man Wikipedia zitieren?
  - Ja, aber nicht dauernd

Hinweise zur Ausarbeitung –2-

- Gezielt und sachlich schreiben
  - Ausführungen zur „Philosophische Überlegungen zu Vorzügen probabilistischer Verfahren im Vergleich zu Dempster‘s Theory of Evidence“ oder zur „Anmerkungen zur Trivialisierung des politischen Diskurs für soziale Netzwerke unter besonderer Berücksichtigung von Twitter“ möglichst kurz halten
  - Füllwörter vermeiden (dabei, hierbei, dann, …)
  - Knappe Darlegung, präzise Sprache
- Eine gute Gliederung ist die halbe Miete
- Kommen Sie zu Aussagen
  - Vorteile, Nachteile, verwandte Arbeiten, mögliche Erweiterungen, Anwendbarkeit, eigene Erfahrungen, …