KASSANDRA

Suchmaschine mit Latent Semantic Indexing

Jette Klein-Berning, Johanna Geiß, Flo Fürstenberger

Irgendwas läuft hier falsch.

- Herkömmliche Suchmaschinen: entweder ein Dokument enthält das Suchwort oder nicht.
- Relevante Dokumente, die das Suchwort nicht enthalten, können nicht gefunden werden.

LSI to the rescue.

- Wunschtraum I: Eine Suchmaschine, die inhaltlich verwandte Texte unabhängig von Einzelwörtern erkennen kann.
- Wunschtraum II: Das ganze bitte ohne langwieriges und mühseliges NLP.

Was ist LSI?

- Latent semantic indexing berücksichtigt nicht nur die Wörter, die ein Dokument tatsächlich enthält, sondern auch die, die es (statistisch gesehen) seinem Inhalt nach enthalten könnte.
- Hierbei verkleinert es angenehmerweise auch den Vektorraum der Term-Dokument-Matrix von etlichen 100.000 auf ca. 500 Dimensionen.

Was ist LSI?

- LSI berechnet die Ähnlichkeit von Dokumenten, ohne auch nur die geringste Ahnung von deren Inhalt zu haben, es beruht ausschließlich auf Matrixalgebra.

Zielsetzung

- Implementierung eines Systems aus 3 Komponenten:
  - Harvester
  - Indexer
  - Web Frontend

Harvester

- Einlesen der Daten via HTTP (ermöglicht Berücksichtigung von .htaccess) oder aus lokalen Dateien.
- Konvertieren der Daten zu unformatiertem Text mit externen Hilfsprogrammen (pdf2txt, ps2txt, rtf2txt, etc.)

Harvester

Beispieltext UniMut

- Entfernen der Formatierung

Unis lernen lesen (5.1.2003)

Für die einen ist 2003 das Jahr der Bibel, für Baden-Württemberg soll 2003 das Jahr der Auslese werden. Glaubt man Minister Frankenberg, wird der bereits von Minister Trotha seit Jahren verkündete und bisher durch drei Stufen vorbereitete Hochschulinnovationsschubdurchbruch jetzt endlich kommen -- und zwar weder durch Studiengebühren noch durch neue Studiengänge, weder durch mehr Geld für die Hochschulen noch durch Orientierungsprüfungen, sondern durch Auswahlverfahren. Hochschulen und Studierende werden -- so Stuttgart -- gleichermaßen profitieren, wenn in Zukunft anstelle von Abinoten, Eignung und Motivation entscheiden. Ermittelt werden Eignung und Motivation über Noten. "Kriterien hierbei sind die Leistungen in den Kernfächern Deutsch, einer Fremdsprache und Mathematik sowie Noten in den Fächern, die besondere Aussagen darüber zulassen, ob der Bewerber für den gewählten Studiengang geeignet ist." De facto findet also eine Veränderung des bisherigen Verfahrens statt. Wer zuvor den Abischnitt durch geschicktes Belegen noch ein bisschen nachbessern konnte, weil er oder sie zwar vielleicht hochmotiviert für Jura oder Sport war, aber in Mathe oder Französisch nicht so gut, muss in Zukunft anders rechnen. Welches die Fächer sind, die Aussagen über die Eignung zulassen, entscheiden die Hochschulen. So wird in Zukunft voraussichtlich für Jura in Heidelberg die Lateinnote mit ausschlaggebend sein. Wer Latein abwählt und stattdessen vielleicht Informatik macht, um sich hier als JuristIn später zu spezialisieren, sollte sich eine andere Hochschule suchen - fürs Abwählen können auch Maluspunkte vergeben werden...

Harvester

- Entfernen von Stoppwörtern

Unis lernen lesen (512003) Für die einen ist 2003 das Jahr der Bibel für Baden-Württemberg soll 2003 das Jahr der Auslese werden Glaubt man Minister Frankenberg wird der bereits von Minister Trotha seit Jahren verkündete und bisher durch drei Stufen vorbereitete Hochschulinnovationsschubdurchbruch jetzt endlich kommen und zwar weder durch Studiengebühren noch durch neue Studiengänge weder durch mehr Geld für die Hochschulen noch durch Orientierungsprüfungen sondern durch Auswahlverfahren Hochschulen und Studierende werden so Stuttgart gleichermaßen profitieren, wenn in Zukunft anstelle von Abinoten Eignung und Motivation entscheiden Ermittelt werden Eignung und Motivation über Noten Kriterien hierbei sind die Leistungen in den Kernfächern Deutsch einer Fremdsprache und Mathematik sowie Noten in den Fächern die besondere Aussagen darüber zulassen ob der Bewerber für den gewählten Studiengang geeignet ist De facto findet also eine Veränderung des bisherigen Verfahrens statt Wer zuvor den Abischnitt durch geschicktes Belegen noch ein bisschen nachbessern konnte weil er oder sie zwar vielleicht hochmotiviert für Jura oder Sport war aber in Mathe oder Französisch nicht so gut muss in Zukunft anders rechnen Welches die Fächer sind die Aussagen über die Eignung zulassen entscheiden die Hochschulen So wird in Zukunft voraussichtlich für Jura in Heidelberg die Lateinnote mit ausschlaggebend sein Wer Latein abwählt und stattdessen vielleicht Informatik macht um sich hier als JuristIn später zu spezialisieren sollte sich eine andere Hochschule suchen fürs Abwählen können auch Maluspunkte vergeben werden

Harvester

- Entfernen von Wörtern, die in jedem oder nur einem Text vorkommen

Abinote
Abischnitt
Abwählen
Ausgrabung (8x)
Aussage (2x)
Auswahlverfahren
Belegen
Bewerber
Bibel
Baden-Württemberg
Deutsch
Eignung (3x)
Fach
Frankenberg
Französisch
Fremdsprache
Fürstenfeldberg
Heidelberg
Getüm (4x)
Hochschulinnovationsschub-durchbruch
Informatik
Jahr (3x)
Jura (2x)
JuristIn
Kernfach
Kriterium
Latein
Lateinnote
Leistung
Maluspunkte
Mathematik (2x)
Minister (2x)
Motivation (2x)
Noten (2x)
Orientierungsprüfung
Sport
Studiengang (2x)
Studiengebühr
Studierende
Stufe
Stuttgart
Trotha
Uni
Veränderung
Verfahren
Zukunft (3x)

Harvester

- Schreiben der Daten in eine PostgreSQL-Datenbank

	d1	d2	d3	d4	d5
Abinote	1	0	0	1	0
Abischnitt	1	1	0	1	0
Abwählen	1	0	1	1	0
Ausgrabung	8	0	1	0	1
Aussage	2	1	1	0	1
Auswahlverfahren	1	1	1	1	0
Belegen	1	0	0	1	1
Bewerber	1	0	0	0	1
Bibel	1	1	0	0	0
Baden-Württemberg	1	0	1	1	1
Deutsch	1	0	1	0	1
Eignung	3	0	0	1	0
Fach	1	1	0	1	0
Frankenberg	1	1	1	0	1
Französisch	1	0	0	0	1
Fürstenfeldberg	1	0	0	1	0
Geld	1	1	0	0	0
Getüm	4	0	1	0	0
…	…	…	…	…	…

Indexer

- Laden der Term-Dokument-Matrix aus der Datenbank.

	d1	d2	d3	d4	d5
Abinote	1	0	0	1	0
Abischnitt	1	1	0	1	0
Abwählen	1	0	1	1	0
Ausgrabung	8	0	1	0	1
Aussage	2	1	1	0	1
Auswahlverfahren	1	1	1	1	0
…	…	…	…	…	…

Indexer

- Normalisierung der Termgewichtung
  - lokal (logarithmic local weighting)
  - global (inverse document frequency)

	d1	d2	d3	d4	d5
Abinote	1,44	0	0	2,83	0
Abischnitt	2,47	3,51	0	1,85	0
Abwählen	3,03	0	1,72	2,87	0
Ausgrabung	4,47	0	2,78	0	4,47
Aussage	1,62	2.23	0,76	0	1,62
Auswahlverfahren	6,22	7,24	3,45	9,23	0
…	…	…	…	…	…

Indexer

- Anwenden von Singular Value Decomposition auf die Term-Dokumentmatrix (SVDPACKC, Perl Modul LSI)

SVD methods are based on the following theorem of linear algebra, whose proof is beyond our scope: Any $M \times N$ matrix **A** whose number of rows M is greater than or equal to its number of columns N, can be written as the product of an $M \times N$ column-orthogonal matrix **U**, an $N \times N$ diagonal matrix **W** with positive or zero elements (the *singular values*), and the transpose of an $N \times N$ orthogonal matrix **V**. The various shapes of these matrices will be made clearer by the following tableau:

$$\begin{pmatrix} & \\ & \\ & \\ & \\ & \\ & \end{pmatrix} = \begin{pmatrix} & \\ & \\ & \\ & \\ & \\ & \end{pmatrix} \cdot \begin{pmatrix} w_1 & & & \\ & w_2 & & \\ & & \cdots & \\ & & \cdots & \\ & & & w_N \end{pmatrix} \cdot \begin{pmatrix} & \\ & \\ & \end{pmatrix} \tag{2.6.1}$$

The matrices **U** and **V** are each orthogonal in the sense that their columns are orthonormal,

$$\sum_{i=1}^{M} U_{ik} U_{in} = \delta_{kn} \qquad \begin{matrix} 1 \le k \le N \\ 1 \le n \le N \end{matrix} \tag{2.6.2}$$

$$\sum_{j=1}^{N} V_{jk} V_{jn} = \delta_{kn} \qquad \begin{matrix} 1 \le k \le N \\ 1 \le n \le N \end{matrix} \tag{2.6.3}$$

Indexer

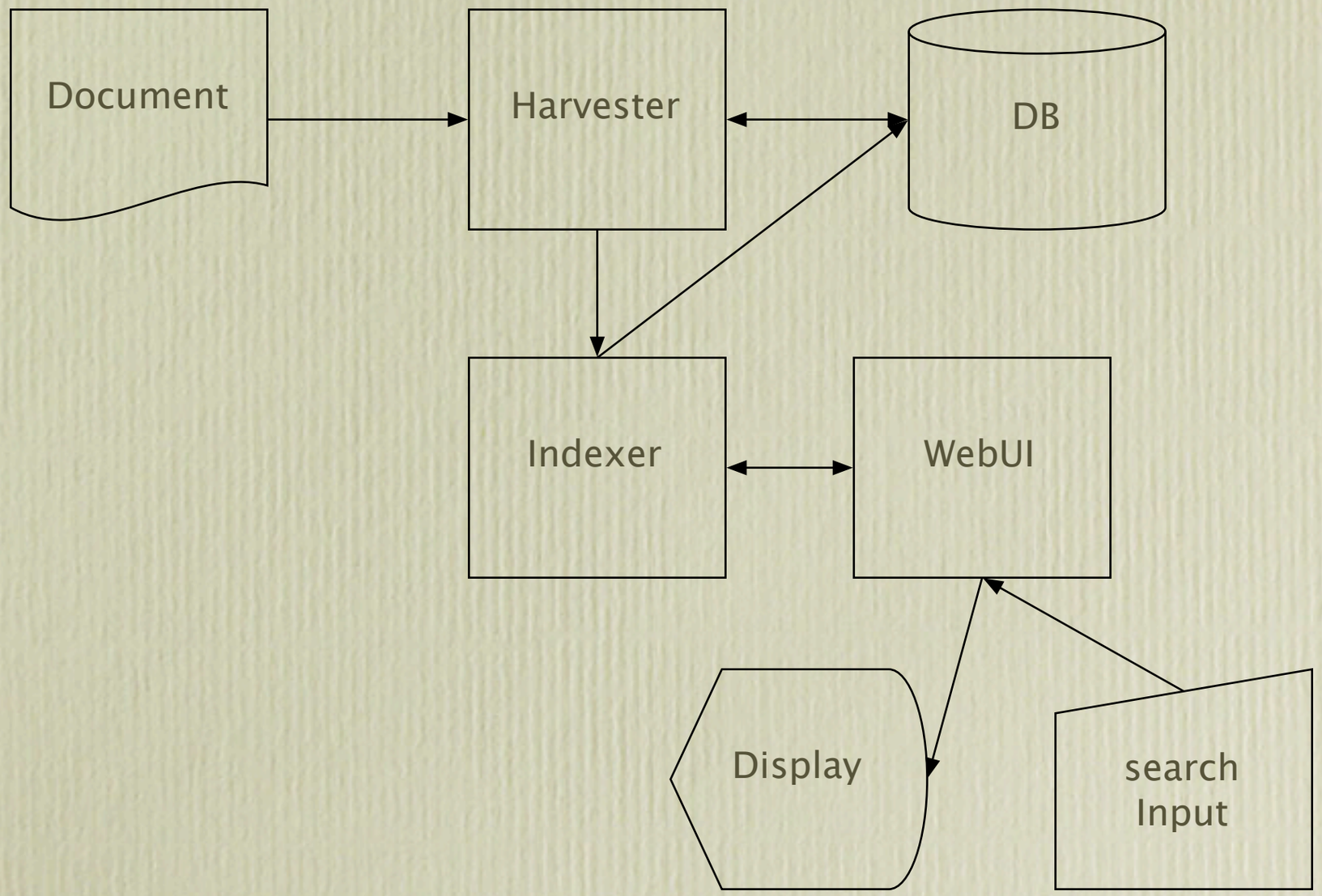
- Schreiben der Matrix in die Datenbank

	d1	d2	d3	d4	d5
Abinote	1,44	0	0	2,83	0
Abischnitt	2,47	3,51	0	1,85	0
Abwählen	3,03	0	1,72	2,87	0
Ausgrabung	4,47	0	2,78	0	4,47
Aussage	1,62	2.23	0,76	0	1,62
Auswahlverfahren	6,22	7,24	3,45	9,23	0
…	…	…	…	…	…

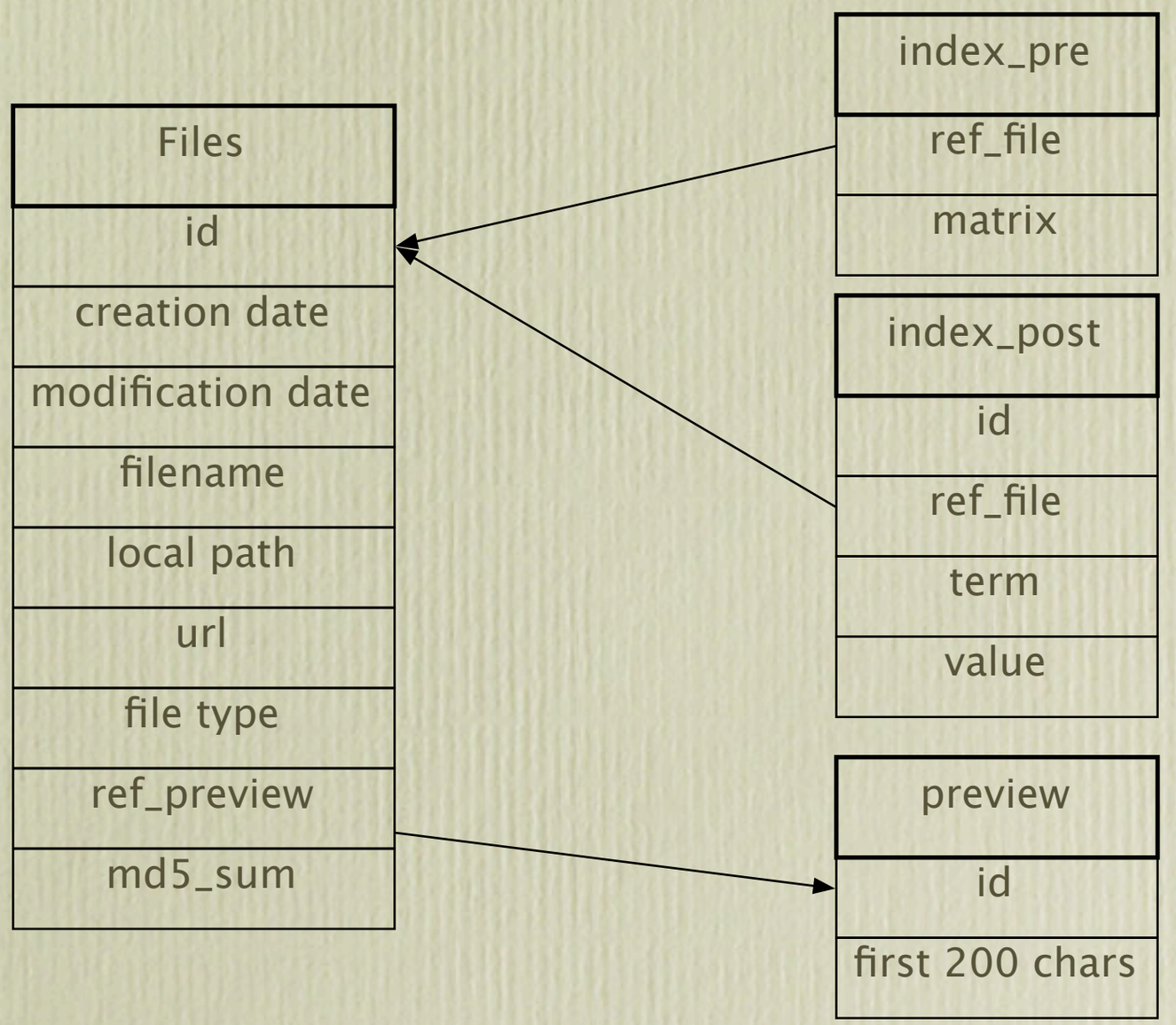
Web Frontend

- Eingabemaske für Suchstring
- Suchergebnisse werden nach Relevanz sortiert ausgegeben.
- Ausgabe der ersten 200 Zeichen jedes gefundenen Dokuments.

Programmstruktur



Datenstruktur



KASSANDRA

Suchmaschine mit Latent Semantic Indexing

Jette Klein-Berning, Johanna Geiß, Flo Fürstenberger