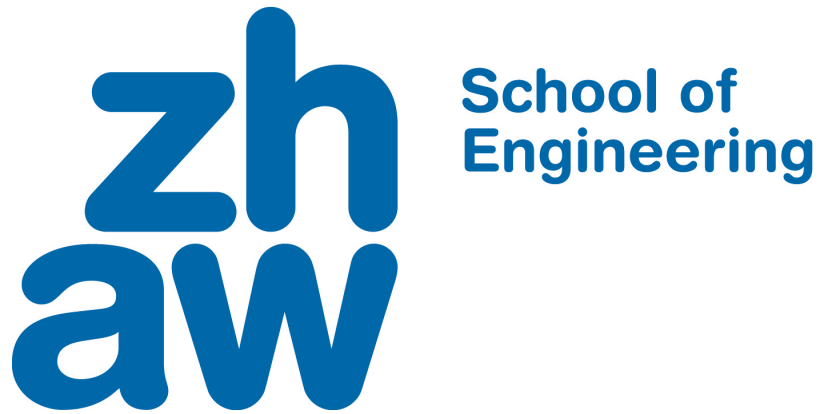


Zürcher Hochschule  
für Angewandte Wissenschaften



## INFORMATION RETRIEVAL

### Seminararbeit FS 2014

Student: Micha Schönenberger

Dozent: Dr. Ruxandra Domenig

© 2014

# Inhaltsverzeichnis

<b>Abbildungsverzeichnis</b>	<b>III</b>
<b>Tabellenverzeichnis</b>	<b>IV</b>
<b>Verzeichnis der Listings</b>	<b>V</b>
<b>1. Versionierung</b>	<b>1</b>
<b>2. Aufwände</b>	<b>2</b>
<b>3. Einleitung</b>	<b>3</b>
3.1. Das Projekt . . . . .	3
3.1.1. Ausgangslage . . . . .	3
3.1.2. Ziel der Arbeit . . . . .	3
3.1.3. Aufgabenstellung . . . . .	4
3.1.4. Erwartetes Resultat . . . . .	5
3.1.5. Geplante Termine . . . . .	5
<b>4. Die Search-Engine</b>	<b>6</b>
4.1. eingesetzte Software . . . . .	6
4.2. eingesetzte Hardware . . . . .	7
<b>5. Vorbereitung für Vergleich</b>	<b>8</b>
5.1. Windows Index Suche . . . . .	8
5.2. Mac OS X Spotlight . . . . .	10
5.3. Syntax für Suche mit Lucene / Luke . . . . .	11
5.3.1. Felder . . . . .	12
5.3.2. Wildcard Suche . . . . .	12
5.3.3. Fuzzy Suche . . . . .	12
5.3.4. Proximity Suche (Nachbarschaftliche) . . . . .	12
5.3.5. Range Suche (Bereichssuche) . . . . .	13
5.3.6. Boosting Suche (verstärktes Suchen) . . . . .	13
5.3.7. Boolesches Suchen . . . . .	13
<b>6. Vergleich der Ergebnisse</b>	<b>14</b>
<b>7. Fazit</b>	<b>15</b>

---

<b>A. Anhang</b>	<b>i</b>
<b>Literaturverzeichnis</b>	<b>ii</b>

## Abbildungsverzeichnis

5.1. Windows Indexing - Default Einstellungen . . . . .	9
5.2. Mac OS X Spotlight - Default Einstellungen . . . . .	10
5.3. Such-Maske von Luke . . . . .	11

## Tabellenverzeichnis

1.1. Versionierung Dokumentation . . . . .	1
2.1. Aufwände Seminararbeit . . . . .	2
3.1. geplante Termine . . . . .	5

## Verzeichnis der Listings

## 1. Versionierung

Version	Datum	Beschreibung
V0.1	18.03.2014	Ersterstellung Dokument
V0.2	11.05.2014	Einarbeitung in Lucene, Installation Software, Lesen LiA
V0.3	13.05.2014	Buch LiA lesen / Coden
V0.4	14.05.2014	Doku schreiben, Query Optionen studieren
V0.5	15.05.2014	Doku LATEX, Query Optionen studieren
V0.6	17.05.2014	eingesetzte Hardware, Kapitel Vorbereitung für Vergleich
V0.8	18.05.2014	Kapitel Vorbereitung für Vergleich, Code Refactoring

Tabelle 1.1.: Versionierung Dokumentation

## 2. Aufwände

Datum	Zeit	Beschreibung
18.03.2014	1.25h	Ersterstellung Dokument -Download und Installation Lucene -Erste Versuche mit Lucene
26.03.2014	3.75h	Suche nach Thema für Eingabe EBS -Online Suche Funktionalitäten Lucene
11.05.2014	4.25h	Installation von Netbeans -Update Eclipse -Lucene in Action: Einlesen Kapitel 1 -1. Versuch Indexer und Searcher (gemäss Beispielen LiA)
12.05.2014	1.75h	siehe 11.05.2014
13.05.2014	7h	-Indexierung erweitern auf alle Subdirectories des Root-Folders -Einschränken von File-Extensions: momentan nur Textdateien (.txt, .h, .c, .java)
13.05.2014	2.75h	-Dokumentation Kapitel 3.1.3 bis 3.1.4 -Quellenverzeichnis erstellen -Kapitel 4.1 eingesetzte Software -Kapitel 5.3 Syntax für Suche mit Lucene / Luke -Einarbeiten in Query mit Lucene/Luke
15.05.2014	4.25h	-Dokumentation von Word in LATEX umschreiben -Kapitel 5.3 Syntax für Suche mit Lucene / Luke
17.05.2014	2.25h	-Kapitel 5 mit teils Unterkapiteln -Kapitel 4.2 eingesetzte Hardware
18.05.2014	2.25h	-Kapitel 5 mit teils Unterkapiteln -Code Refactoring and cleaning

Tabelle 2.1.: Aufwände Seminararbeit



## 3. Einleitung

### 3.1. Das Projekt

#### 3.1.1. Ausgangslage

In den knapp 4 Jahren an der ZHAW haben sich verschiedenste Dateien angehäuft. Die Thematiken überschneiden die Module und sind somit nicht immer über eine übersichtliche Ordnerstruktur auffindbar.

Die vielen Typen von Dokumenten (Office, pdf, txt, java...) vorhanden sind, wird durch die Betriebssystem integrierte Suche nicht immer das erwartete Resultat geliefert. Alle diese Dokumente sollen über Lucene mit einem Index versehen und für eine effiziente Suche optimiert werden. So können unter anderem auch PDF inhaltlich indexiert werden, was bei der Search-Engine des Betriebssystems nicht funktioniert.

#### 3.1.2. Ziel der Arbeit

Das Ziel der Arbeit soll ein direkter Vergleich zwischen der Suchresultate von Lucene mit denjenigen des Betriebssystems stattfinden.

Aufgrund der subjektiven Sicht der suchenden Person soll schlussendlich entschieden werden, welche Search-Engine die besseren Suchresultate liefert. Es sollen mindestens zwei reale Begriffe gesucht werden und anhand dieser ein Fazit gezogen werden (eventuell mit Verbesserungsvorschlägen für die Optimierung von Lucene).

Für die Indexierung von PDF, welche nicht durch die Search-Engine des Betriebssystems bei der Indexierung eingeschlossen werden soll Lucene Abhilfe schaffen.

### 3.1.3. Aufgabenstellung

1. Definierung der Suchkriterien über alle zu indexierenden Dateien
2. Erstellen einer einfachen Demo-Applikation in Java
3. Indexierung aller Schuldateien
4. Implementierung für die Indexierung von PDF. Wenn dies nicht direkt möglich sein sollte in Lucene, werden PDF in Textdateien umgewandelt und dann indexiert
5. Überprüfung der Suchresultate und Optimierung von Lucene
6. Reale Suche von mindestens zwei Begriffen in Lucene und direkter Vergleich mit den Suchresultaten des Betriebssystems
7. Fazit der Implementierung von Lucene. Sind die Suchresultate besser als diejenigen des Betriebssystems? Begründung

#### 3.1.4. Erwartetes Resultat

1. PDF sollen in den Suchresultaten mitberücksichtigt werden
2. Die Eingabe der Suchbegriffe soll intuitiv und einfach gehalten werden
3. Die Qualität der Suchresultate von Lucene soll diejenige vom Betriebssystem übertreffen. Falls dies nicht der Fall sein sollte, soll eine kurze Analyse aufzeigen, wieso das Betriebssystem bessere Resultate hervorbringen kann

#### 3.1.5. Geplante Termine

Datum	Beschreibung
14.03.2014	Kick-Off Meeting
11.06.2014	Abgabe der schriftlichen Arbeit (1 Woche vor Präsentation)
18.06.2014	Präsentation der Arbeit

Tabelle 3.1.: geplante Termine

## 4. Die Search-Engine

### 4.1. eingesetzte Software

Um mit Lucene programmieren zu können, wurden folgende Software eingesetzt:

- Mac OS X 10.9 (Mavericks)
- Eclipse SDK (Indigo), Version 3.7.2  
<http://www.eclipse.org/downloads/packages/release/indigo/sr2>
- Java(TM) SE Runtime Environment (build 1.8.0\_05-b13)  
<http://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html>
- Lucene Version 3.0.2 (für Indexierung)  
<http://www.apache.org/dyn/closer.cgi/lucene/java/3.0.2>
- Apache PDFBox Version 1.8.5 (Open Source Library für Indexierung von PDF)  
<http://pdfbox.apache.org/downloads.html>
- Luke - Lucene Index Toolbox - v 3.5.0 (2011-12-28)  
stellt Indexierte Dateien von Lucene dar und erlaubt die Suche nach Stichwörtern.  
<https://code.google.com/p/luke/downloads/detail?name=lukeall-3.5.0.jar&>
- Tika  
<http://www.apache.org/dyn/closer.cgi/tika/tika-app-1.5.jar>

## 4.2. eingesetzte Hardware

- Apple Mac Book Air 2011 (MacBookAir4,2)
  - OS: MAC OS X 10.9.2
  - RAM: 4GB 1333 MHz DDR3
  - CPU: 1.8 GHz (2677M) Dual-Core i7 mit 4 MB on-chip L3 cache
  - Harddisk: SSD 256GB - APPLE SSD SM256C Media
- Asus P7H55E
  - OS: Windows 8.1 Professional
  - RAM: 8GB DDR3
  - CPU: Intel Core i5 661 @ 3.33GHz
  - Harddisk: Western Digital black Edition - 1TB (WD 1002FAEX-00Y9A0)

## 5. Vorbereitung für Vergleich

Bevor ein Vergleich zwischen den Verschiedenen Search-Engines gemacht werden kann, sollte ein Überblick über die Search-Engines erstellt werden.

### 5.1. Windows Index Suche

Standardmässig werden bei Windows nur einzelne Ordner indexiert. Gleichzeitig indexiert Windows nicht die kompletten Datei Inhalte, sondern lediglich die Einstellungen.

Um die Inhalte auch zu indexieren, muss dies in den Einstellungen geändert werden (siehe Abbildung 5.1 rechte Seite).

Um nun ein Netzlaufwerk indexieren zu können (was hier auch gewünscht ist), muss ein Add-in von Microsoft installiert werden. Dieses nennt sich «Windows Desktop Search: Add-in for Files on Microsoft Networks» und ist auf der Microsoft Homepage verfügbar<sup>1</sup>.

Da dies leider nach einem Neustart nicht funktionierte, wurde eine weitere Möglichkeit gesucht.

So wurde gemäss dem Microsoft Technet-Artikel<sup>2</sup> folgende Punkte erledigt:

- Erstellen einer neuen Windows Library «ZHAW Docs» erstellt.
- Erstellen eines neuen Ordners C:\ZHAW Docs
- Hinzufügen des neu erstellten Ordners zur neuen Library
- Ordner C:\ZHAW Docs löschen
- Über cmd (als Administrator) folgenden Befehl ausführen:  
mklink /d "C:\ZHAW Docs" "\\<IP Synology>\<Share-Path>"  
Mit diesem Befehl wurde ein symbolischer Link des Netzwerkpfades auf den lokalen Pfad gesetzt.

---

<sup>1</sup><http://www.microsoft.com/en-us/download/details.aspx?id=3383>

<sup>2</sup><http://social.technet.microsoft.com/Forums/windows/en-US/afb904c1-1c61-4aae-b6b1-5cf525b9f8de/how-do-i-get-windows-7-to-index-a-network-mapped-drive?forum=w7itpronetworking>

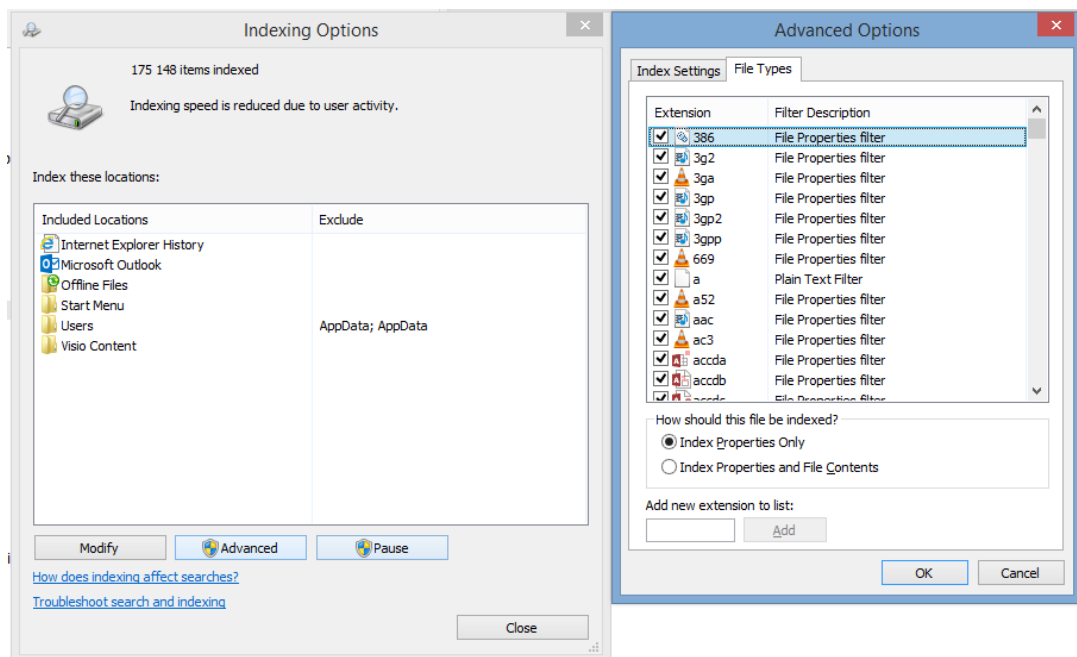


Abbildung 5.1.: Windows Indexing - Default Einstellungen  
Quelle: eigener Screenshot

## 5.2. Mac OS X Spotlight

Beim MAC OS X Spotlight gibt es weniger Einstellungen, die man selber vornehmen kann. So kann zum Beispiel nicht gewählt werden, ob das ganze File oder nur die Einstellungen oder Metadaten zum Index hinzugefügt werden.

Abbildung 5.2 zeigt die Einstellungsmöglichkeiten. Wird hier Nummer 7 (=PDF) ausgewählt, indexiert MAC OS X automatisch den ganzen Inhalt des PDF, vorausgesetzt der Text im PDF ist lesbar.

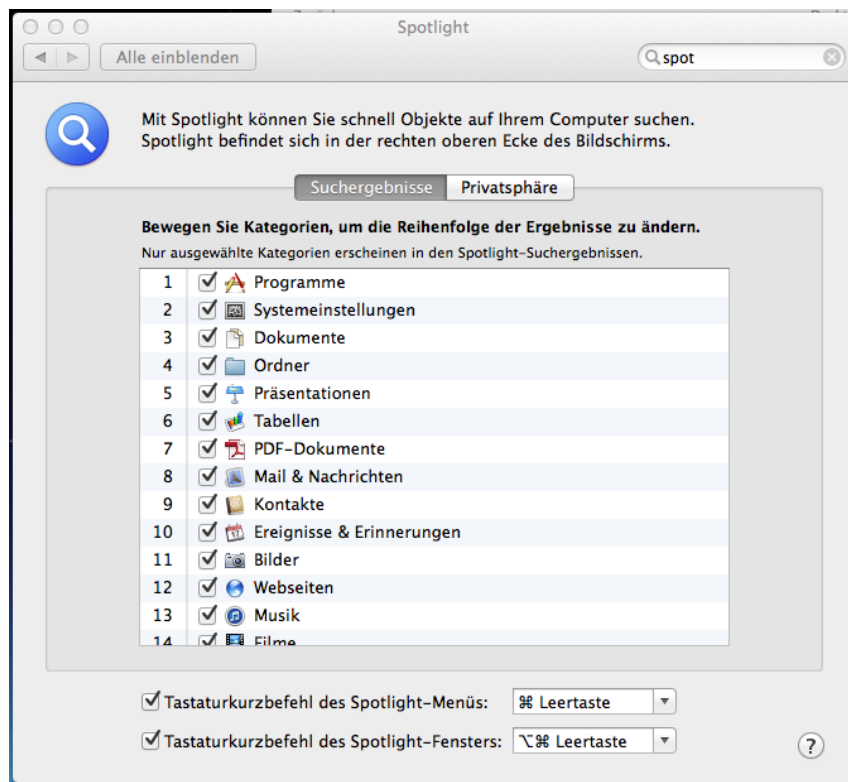


Abbildung 5.2.: Mac OS X Spotlight - Default Einstellungen  
Quelle: eigener Screenshot

Beim MAC OS ist war es nicht notwendig, ein Netzlaufwerk zu indexieren, da die Daten bereits lokal auf der Festplatte vorhanden waren.



### 5.3. Syntax für Suche mit Lucene / Luke

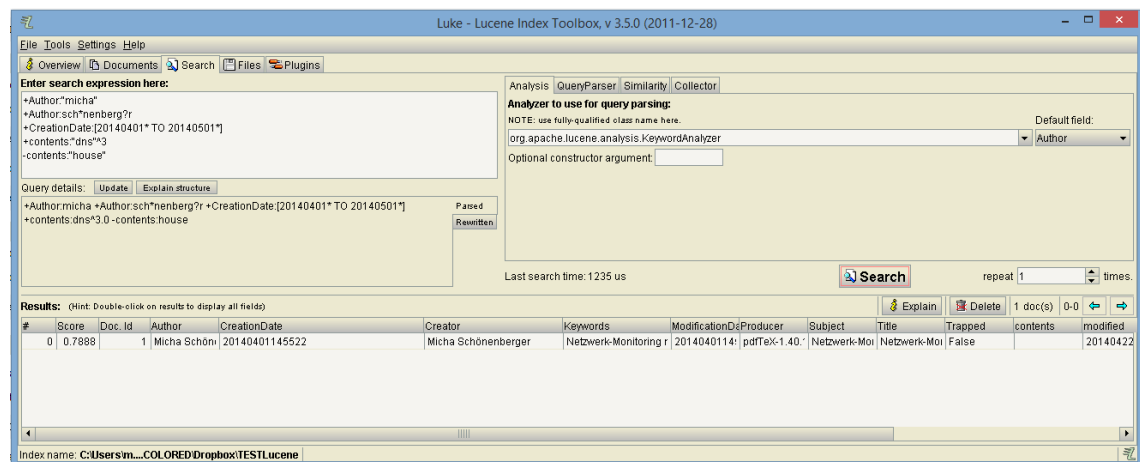


Abbildung 5.3.: Such-Maske von Luke  
Quelle: eigener Screenshot

Anbei einige Beispiele, wie man den Syntax optimal einsetzen kann für die Suche mit Luke/Lucene:

### 5.3.1. Felder

contents:prtg	sucht das Wort "prtg" im Feld contents
Author:micha monitor	sucht das Wort "micha" im Feld Author, das Wort "netzwerk" im default Feld
contents:?mit prtg?	sucht die Wörter "mit prtg" als ganzes im Feld contents

### 5.3.2. Wildcard Suche

pr?g	Das ? ersetzt einen einzelnen Charakter. So wird hier z.B. nach ?prtg?, aber auch nach ?prag? gesucht
bildschirm*	* steht für beliebig viele Zeichen (0 bis x) . Suche nach z.B. "bildschirmeingabe" oder "bildschirme"
ver*en	* steht auch hier für beliebige Zeichen (0 bis x) Suche nach z.B. "verwerfen" "verwalten" oder "veruntreuen"

Grundsätzlich gilt: \* und ? dürfen nicht am Anfang der suche sein. (bei Luke gibt es eine Einstellung, mit der man \* am Anfang erlauben kann)

### 5.3.3. Fuzzy Suche

Die Fuzzy Suche basiert auf dem Levenshtein Distanz Algorithmus<sup>3</sup>, auf welchen hier nicht eingegangen wird.

Um eine Fuzzy Suche durchzuführen, wird mit der Tilde ( ) gearbeitet.

roam	Sucht ähnliche Wörter zu roam
roam 0.8	Gibt die gesuchte Ähnlichkeit an [0-1]. Je grösser die Zahl, desto ähnlicher.

### 5.3.4. Proximity Suche (Nachbarschaftliche)

Lucene erlaubt es, zwei Wörter in einer bestimmten Distanz zu suchen.

"prtg monitor"~10	Sucht die Wörter "prtg" und "monitor" mit einem maximalen Abstand von 10 Wörtern
-------------------	---

<sup>3</sup>siehe <http://www.levenshtein.de/>

### 5.3.5. Range Suche (Bereichssuche)

Lucene erlaubt es, zum Beispiel Daten in einem bestimmten Bereich zu suchen.

modified:[20140501 TO 20140515]	Sucht eine Datei mit einem Änderungsdatum zwischen dem 01.05.2014 und 15.05.2014
Title:{Anton TO Max}	Sucht eine Datei mit dem Titel zwischen Anton und Max (alphabetisch)

### 5.3.6. Boosting Suche (verstärktes Suchen)

Lucene erlaubt es, gewissen Suchbegriffen eine stärkere Gewichtung zu verleihen. Durch diesen Boost wird das Ergebnis die Relevanz der gefundenen Dateien verändert.

prt看^ 4 monitoring	Sucht eine Datei mit «prt看» und «monitoring» wobei «prt看» 4-fach gewertet wird.
--------------------	---

### 5.3.7. Boolesches Suchen

Ganz wichtig beim Suchen sind die verschiedenen Verknüpfungen der eingegebenen Suchbegriffe. Erlaubt sind AND, +, NOT, -

prt看 OR monitor	Sucht entweder nach «prt看» oder nach «monitor»
prt看 AND monitor	Sucht nach «prt看» und nach «monitor». Beide Wörter müssen im selben Dokument enthalten sein.
prt看 NOT monitor	Sucht ein Dokument, in welchem «prt看» vorkommt, «monitor» aber nicht.
NOT prt看	NOT darf nicht mit nur einem Wort verwendet werden. Das Resultat wird hier leer sein.

Anstelle von AND kann auch && oder + verwendet werden.

Anstelle von NOT kann auch ! oder - verwendet werden.

Quelle Kapitel 5.3: [\[1\]](#)

## 6. Vergleich der Ergebnisse

Wie in der Aufgabenstellung erwähnt, sollen die Suchresultate von Lucene mit denjenigen des Betriebssystems verglichen werden. Die Suche im Betriebssystem ist sehr intuitiv. Es wird der gewünschte Suchbegriff eingegeben und nach ein paar Sekunden (oder Minuten), je nach dem ob alle Ordner indexiert waren oder nicht, erscheint das Resultat.

Bevor jedoch ein Vergleich stattfinden kann, wird aufgezeigt, wie mit Luke eine effiziente Suche ausgeführt werden kann über die indexierten Dateien der Lucene-Applikation. Diese werden im nächsten Unterkapitel erläutert:

## 7. Fazit

## A. Anhang

## Literaturverzeichnis

- [1] *Lucene Query Parser Syntax*. [http://lucene.apache.org/core/2\\_9\\_4/queryparsersyntax.html](http://lucene.apache.org/core/2_9_4/queryparsersyntax.html), may 2014. 5.3.7
- [2] MCCANDLESS, MICHAEL; HATCHER, ERIK; GOSPONDENETIC OTIS: *Lucene in Action - Second Edition*. Manning Publications Co, 2010.