Rechercheberichт

EMM16
André Benda

Inhalt

1. Begriffe

1.1. Nutzer/Zielgruppe

Die Applikation wird von Studienberatungsseiten in ihre Websites implementiert. Darauf greifen Studieninteressierte zu, um sich über Module und Kurse an Hochschulen zu informieren.

1.2. Webadministratoren

Sind Personen, die für Webauftritte verantwortlich sind. Diese binden unsere Applikation ein und erwarten einen möglichst geringen Wartungsaufwand.

1.3. Administration

In einem Backend muss der Webadministrator der Website die Quellen für den Import der Daten angeben. Dabei sollen sowohl OPAL und Moddle unterstützt werden.

1.4. Webserver

Dieser liefert grundsätzlich Webseiten aus. Dabei laufen im Hintergund noch viele andere Anwendungen, die für unser Projekt von Nöten sind. So z.B. ein PHP-Interpreter, ein Dateisystem und ein Cron-Manager (dieser führt Aufgaben regelmäßig ohne zutun aus).[i]

1.5. Framework

Ist ein Programmiergerüst, dass wiederkehrende Aufgaben erleichtert. Es dient sozusagen als Rahmen für Applikationen.

1.6. Kurs und Metadaten

Kurse sind zusätzliche Angebote die Online zusätzliche Informationen bereitstellen. In unserem Fall sind Kurse zu Modulen begleitende Angebote zu Modulen, die eine Hochschule anbietet. Meta-Daten sind, die zu einem Objekt (Kurs) existieren und diesen beschreiben Attribute. Oft folgen sie einem bestimmten Muster, sodass Daten geordnet und vergleichbar sind.

1.7. E-Learning Portal

Sind Online Plattformen, die zur Unterstützung der Lehre erstellt werden und meist passend zu den Kursangeboten der Hochschulen sind. Sie beinhalten sowohl Dokumente, die für die Studierenden notwendig sind als auch Metadaten, die für unsere Suchmaschine genutzt werden können.

1.8. Crawler

Sind Programme, die wiederkehrend Daten von Websites einsammeln und die Daten an den zugehörigen Speicher übergeben. Im vorliegenden Fall sammelt der Crawler die öffentlichen Kursdaten der E-Learning Portale ein.

1.9. Inverted Index

Hierbei werden alle möglichen Suchbegriffe (reduzierte Tokens) mit Zeigern auf die Fundstellen abgespeichert, zeigt ein Begriff auf mehrere Dokumente, so existiert auch nur ein Datensatz. Dadurch wird auch die Verknüpfung verschiedener Querys ermöglicht.[ii]

1.10. Dokumente

Dokumente sind das äquivalent zu Datensätzen in Indextrukturen. Sie beinhalten die Daten mit verschiedenen Parametern.

1.11. Tokenizer und Stemming

Ein Tokenizer zerlegt gegebene Texte in einzelne Tokens (in der Deutschen Sprache Wörter und Phrasen) und entfernt Apostrophe und Satzzeichen. In Lucene heißen Tokenizer Analyzer. Für die deutsche Sprache gibt es den GermanAnalyzer, der zusätzlich Sprachgegebenheiten und Stoppwörter (z.B. und, oder, an, zu) entfernt, die bei der Suche keine Rolle spielen.

Stemming reduziert Wörter auf ihre Stammformen, um den Index zu verkleinern und die Anfrageperformance zu optimieren. Davon abzugrenzen sind Wörterbuchbassierte Manipulationen, die zusätzlich zwischen Wörtern Verbindungen herstellen können und Synonyme hinzufügen.[iii]

1.12. Web-basierte Simulation

Web- basierte Simulation (WBS) ist die Simulation von Server auf Computer. Speziell wird Webseite darauf getestet. Während Entwicklung von Software ist Web- basierte Simulation hilfreich.

1.13. Suchmaschine

Eine Suchmaschine ist hergestellt, um die in einem Computer oder einem Computernetzwerk wie z. B. dem World Wide Web gespeicherte Dokumenten zu recherchieren. Internet Suchmaschinen haben ursprünglich Information-Retrieval-Systemen. Das bedeutet, dass die Suchmaschine einen Schlüsselwort-Index für die Dokumentbasis erstellen, um Suchanfragen über Schlüsselwörter mit einer nach Relevanz geordneten Trefferliste zu beantworten. Die Suchmaschine soll folgende Bestandteile bzw. Aufgabenbereiche erfüllen. Zuerst ist Erstellung und Pflege eines Index (Datenstruktur mit Informationen über Dokumente), zweitens ist Verarbeiten von Suchanfragen (Finden und Ordnen von Ergebnissen) und noch Aufbereitung der Ergebnisse in einer möglichst sinnvollen Form.

1.14. Moodle

Moodle ist ein freies objektorientiertes Kursmanagementsystem und eine Lernplattform. Die Software bietet die Möglichkeiten zur Unterstützung kooperativer Lehr- und Lernmethoden.

1.15. OPAL

OPAL ist eine ganzheitliche und hochschulübergreifende IT-Struktur für E-Learning. Diese Lernplattform wird an 17 sächsischen Bildungseinrichtungen eingesetzt und verfügt über mehr als 80.000 Nutzer. Der Name OPAL steht als Abkürzung für Online-Plattform für Akademisches Lehren und Lernen und soll sowohl Lernenden als auch Lehrenden als Hilfsmittel bei der Bewältigung der ihnen im Rahmen der Ausbildung gestellten Aufgaben fungieren und dabei unterschiedlichste Arbeitsabläufe unterstützen. Die Lernplattform OPAL wird zentral durch die BPS Bildungsportal Sachsen GmbH verwaltet. Der technologische Kern dieser Lernplattform bildete das Open Source Lernmanagement System (LMS) OLAT, welches 1999 an der Universität Zürich entwickelt und seitdem fortlaufend weiterentwickelt wird. Seit 2011 erfolgten schrittweise Weiterentwicklungen vom Open Source Projekt OPLAT weg, die in der Lernplattform OLAT Campus mündeten.

1.16. Usability

Ist ein Maß dafür wie intuitiv und einfach ein Nutzer an die gewünschten Informationen gelangt. Dabei spielen Dinge wie Navigation und Präsentation eine wichtige Rolle.

1.17. Suchindex

Ein Suchindex ist eine Sammlung von Verweisen aus anderswo gespeicherte Daten. Meist ist dies eine Datei, in der Informationen über Dokumente oder Webseiten abgelegt werden. Der Vorteil eines Indexes ist, dass man Daten schnell wiederfinden kann.

1.18. Datenbank

Eine Datenbank ist eine Sammlung gespeicherter operationaler Daten, die von Anwendungssystemen benötigt werden.

1.19. Plugins

(zu deutsch etwa "Erweiterungsmodule") werden zur Erweiterung der Funktionalität von Software verwendet. Die entsprechende Software kann während der Laufzeit passende Plug-ins aufspüren und einbinden. Der Begriff wird teilweise auch als Synonym zu Add-on benutzt.

1.20. Webservice

Webservices ermöglichen es externen Systemen, sich in Moodle anzumelden und Aktionen durchzuführen. Diese Schnittstelle wird in unserem Projekt verwendet um Informationen über die verfügbaren Kurse in Moodle zu bekommen.

1.21.Sicherkeitsschlüssel/Token

Sicherheitsschlüssel ermöglichen es externen Systemen, sicher auf Moodle zuzugreifen. Ein Sicherheitsschlüssel wird bei sicheren RSS-Feeds und bei Webservices verwendet.

2. KONZEPTE

2.1. XML

XML ist eine Auszeichnungssprache und ist die Abkürzung von der Extensible Markup Language (engl. „erweiterbare Auszeichnungssprache“). Die Daten werden in Form von Textdateien hierarchisch strukturierter darstellt. Während Plattform und Implementierung unabhängigen Austausch von Daten zwischen Computersystemen insbesondere über das Internet wird XML benutzt. Hier ist XML erforderlich wegen die gute Zusammenpassung von REST und XML.

2.2. HTML/CSS/Javascript

HTML ist Abkürzung von der Hypertext Markup Language (engl. für Hypertext-Auszeichnungssprache), und HTML ist eine Auszeichnungssprache, die auf Text basiert, um digitaler Dokumente wie Texte mit Hyperlinks, Bildern und anderen Inhalten zu strukturieren. HTML ist die WWW-Sprache, weltweiter Standard, plattformunabhängig, einfach und hat kurze Ladezeit für reine HTML-Seiten (reine Text-Darstellung). CSS sind sogenannte Stylesheets, die zur Gestaltung der Website dienen. Da HTML nur begrenzt dynamisch Inhalte verändern kann ohne die komplette Seite neu zu laden, wird noch Javascript benötigt. Javascript macht genau solche Modifikationen möglich und dient auch zur Übetragung von Daten vom User-Cliet zum Server. Zu Javascript existieren zur Vereinfachung diverse Frameworks z.B: JQuery.

2.3. PHP

Bei PHP handelt es sich um eine serverseitige Programmiersprache. Dies hat den Nachteil, dass das PHP-Dokument bei jedem Aufruf neu interpretiert werden muss, solange kein Byte-Code Cache auf dem Server installiert ist. (Erst ab PHP 5.5 wird die Zend Optimizer+ mitgeliefert).

Der PHP-Code wird auf dem Server interpretiert und danach meist als HTML-Dokument an den Browser übermittelt. Die PHP-Syntax lehnt sich stark an der von C und Pearl an.

Da PHP zu den schwach typisierten Sprachen gehört und der Datentyp automatisch initialisiert wird, muss beim Vergleich von Variablen auf möglicherweise auftretende Fehler bei der Typumwandlung geachtet werden.(Ab PHP 7 wird eine Typsicherheit für skalara Datentypen mitgebracht)

Eine andere Besonderheit an PHP ist, dass Arrays nicht notgedrungen mit den natürlichen Zahlen ab Null indiziert sind. Es können aber „Standard“-Arrays mit Arrays mit individuellen Keys zusammengelegt werden.[iv]

2.4. Apache Lucene

Ist ein effizienter Indexer und Searcher für große Textdatenmengen. Der Index besteht aus einzelnen Documents. Die Suchfunktion beinhaltet sowohl ein Ranking als auch Query Manipulation. Die Software steht als OpenSource auch zur kommerziellen Nutzung bereit und wird auch weit verbreitet eingesetzt (Twitter, Apple Finder). Grundsätzlich wurde Lucene für Java entwickelt läuft aber unter kleinen Performanceeinbußen auch unter PHP mittels des Zend Frameworks.[v]

2.5. Zend Framework 2.0

Ist eine Erweiterung für PHP (>=5.3.3), die auf jedem Webserver lauffähig ist und PHP um diverse Funktionen erweitert. Auf dem Framework können Programme objektorientier

entwickelt werden und viele Funktionen einfacher genutzt werden. Das Framework steht unter der Free BSD Lizenz und ist somit für uns nutzbar.[vi]

2.6. Apache SOLR

Apache SOLR(gesprochen Solar) ist eine auf Java basierende open Source Suchplattform.[vii]

Apache SOLR steht als eigener Lucene Server im Hintergrund, sodass das Zend-Framework nicht genutzt werden muss. Diese Variante der Einbindung des Index ist deutlich performanter im Gegensatz zur Implementierung mittels Zend. SOLR lässt sich als Lucene Backend mit einfachen web-service-calls mittels PHP aufrufen.

2.7. REST-Schnittstelle

Eine REST-Schnittstelle ist ein Angebot eines Servers zur Maschine-zu-Maschine Kommunikation. Dabei werden standardisiert Daten ausgetauscht, die einem definierten Format folgen. Die Daten werden Zustandslos ausgeliefert, sind also weder von Server noch Client abhängig.[viii] Eine Authentifizierung ist möglich. Moddle bietet eine Schnittstelle zum Ausgeben der Kursdaten.[ix] OPAL stellt seine Kursdaten unter einer URL mittels HTTP Protokoll zur Verfügung, auch das ist eine REST-Schnittstelle.[x]

3. Aspekte

Beim Projekt EMM16 geht es um die Durchsuchung von XML-Dokumenten, die von verschiedenen Plattformen vorliegen. Es soll eine Suche entstehen, die passende Treffer zu Suchbegriffen liefert. Der klassische Weg wäre alle Daten zu extrahieren und in eine Datenbank zu übertragen, sodass eine Suchengine zu jedem Datensatz bestimmen kann ob es zur Suche passt oder nicht. Dies ist weder effizient noch skalierbar. Die passende Alterative ist ein Index mit passender Suchfunktion. Apache Lucene ist ein Open Source Projekt, welches genau diese Funktionalitäten zur Verfügung stellt und auch kommerziell für weitaus größere Projekte genutzt wird. Mittels des Zend Frameworks ist Lucene, welches eigentlich für Java entwickelt ist, auch unter PHP nutzbar. Lucene skaliert bis zu 150Gb/h beim Indexieren. Die folgenden Abschnitte beziehen sich jeweils auf diese Lösung.

Abzugrenzen hiervon ist ein zusätzliches Tagging von Inhalten. Also das identifizieren von Wortarten, welche Substantive stehen für Orte, Organisationen und Anderes. Dies ist für das Projekt nicht von Relevanz, da die Suche keine Fragen beantworten soll, sondern die treffendsten Kurse für einen Query liefern soll.

3.1. Systemanforderungen und Sicherheit

Die Applikation soll auf jedem beliebigen Webserver lauffähig sein. Das bedeutet sowohl, dass mit den Ressourcen sparsam umgegangen werden muss als auch keine zusätzlichen Elemente installiert werden sollen. Genutzt werden PHP mit ZEND Erweiterung (die Lucene mitbringt) und das Dateisystem, was auf allen Webservern läuft (Wer nicht das ganze Framework einbinden möchte, kann einfach die Lucene und Exception Ordner herauskopieren). Damit sollte die Applikation auf allen Webservern lauffähig sein und Lucene benötigt nur zum Aufbau des Index ein wenig mehr Rechenleistung, was ein normaler Server bewältigen sollte.

Notwenidg ist auch, die Applikation gegen Angriffe auf den Server abzusichern. Das bedeutet einerseits alle sicherheitsrelevanten Aspekte in PHP zu berücksichtigen, als auch das genutzte Dateisystem vom Rest zu isolieren.[xi]

3.2. Indexierung, Speicherung und Suche

Lucene nutzt einen Inverted Index zur Speicherung der einzelnen Dokumente. Jedes Dokument würde hier für einen Kurs stehen und kann verschiedene Felder enthalten. Mögliche Felder sind beispielsweise Titel, Kursleiter und einer Beschreibung. Felder selbst haben die Möglichkeiten Rohdaten zu speichern, indexiert zu werden und einen Wichtigkeitswert (Boost) zu erhalten. Die Daten für die Felder bestehen aus einfachem Text.

Inhalte der Felder werden durch Tokenizer (bei Lucene Analyzer) zerlegt. Grundsätzlich arbeitet Lucene mit einem Whitespace Analyzer, der bei Whitespace die Daten trennt. Jedoch werden auch Sonder- und Satzzeichen sowie Stoppwörter entfernt (z.B. so, dass, und, oder) die für die spätere Suche keine Rolle spielen. Für viele Sprachen erhält Lucene zusätzlich zusammengehörige Phrasen, die nur gemeinsam Sinn ergeben.

Weiterhin werden Wörter durch Stemmingalgorithmen auf ihre Stammform zurückgeführt, sodass der Index verkleinert wird und die Suchgeschwindigkeit steigt. Nachteil ist dabei, dass manche Wörter den gleichen Stamm haben und doch eine andere Bedeutung in der gebeugten Form. Genauso entstehen Fehler bei unregelmäßigen Wörtern, die nur über ineffektive Wörterbuchvergleiche verhindert werden könnten. Da die Stemmingalgorithmen aber auch auf die Suche angewandt werden, entstehen keine negativen Fehler (nicht gefundene Module) sondern höchstens zusätzliche nicht passende Module, was aber bei Suchen mit mehr als einem Begriff auch nahezu ausgeschlossen ist.

Lucene besitzt keinen eigenen Crawler, sodass hier auf alternative Lösungen zu setzen ist. So könnte Apache Nutch zum Einsatz kommen, da hier ein perfektes Zusammenspiel besteht. Jedoch ist auch eine eigene Lösung denkbar, weil wir ja nur über eine sehr abgesteckte Menge an Daten arbeiten.

Der erzeugte Index wird im Dateisystem erstellt. Dieser kann gelöscht und neu erzeugt werden. Es werden nicht die gesamten Dateien gespeichert, sondern nur das für den Index notwendige und wenn gewollt die Rohtexte. Damit ist der Index etwa 30% der Speichergröße der Ausgangsdaten. Es werden nahezu keine Dateien im RAM gehalten, sodass die Last trotz hoher Performance gering ist. Zu empfehlen ist jedoch ein schneller Plattenzugriff.

Grundlage der Suche ist ein vom User erzeugter Query. Nach langem probieren sind alle Suchmaschinen dazu übergegangen reine UND Verknüpfungen zwischen allen Wörtern des Querys vorzunehmen. Verschiedenste Arbeiten haben auch Zipfs Law belegt, dass besagt User-Querys ein bis zwei Wörter lang sind und kaum zusätzliche Suchfeature enthalten (z.B. OR-Verknüpfung, Dateitypensuche).

Lucene erzeugt aus dem Query einen Lucenequery, mit dem selben Analyzer von der Indexierung. Dieser ist in einem bestimmten Format und kann dem Index übergeben werden, sodass Treffer zurückgegeben werden. Hierbei werden sowohl Verknüpfungen (AND, OR und NOT) sowie Wildcard-Querys unterstützt. Möglich, aber nicht effizient sind Fuzzy-Querys, die auch nach ähnlichen Begriffen suchen oder ein zusätzliches Ranking über die Stellung der einzelnen Begriffe im Dokument.

Die Suche gibt eine Liste von Hits (Treffern) zurück, welche eine sortierte Liste von Dokumenten beinhaltet. Das Ranking basiert auf den Matches mit den Inhalten der Felder und den dazugehörigen Boost-Werten. Aus den zurückgegeben Dokumenten können dann wiederum Informationen ausgelesen werden, wenn sie denn im Index gespeichert worden sind. Sonst ist auch ein Rückbezug auf die Originaldaten möglich.

Weiterhin unterstützt Lucene eine automatische Querykorrektur auf Basis aller indexierten Wörter, es werden also kleine Tippfehler erkannt und korrigiert.[xii]

3.3. Erweiterung der Suche

Als Query Expanding wird der Vorschlag erweiterter Suchbegriffe verstanden, falls die Suche keine oder sehr wenige Suchbegriffe liefert. Dafür wird ein komplexer zusätzlicher Index erzeugt, der Token paarweise rankt. Dies ist jedoch sehr rechen- und speicherintensiv und wird für simple Webserver nicht empfohlen, solange kein eigener Indexierungsserver zur Verfügung steht.

Hingegen ist das Vorschlagen ähnlicher Dokumente auf Basis eines ausgewählten Dokuments auch während der Laufzeit effizient umsetzbar, indem eine neue Suche für das gewählte Dokument durchgeführt wird.

Weiterhin möglich ist die Einschränkung des durchsuchten Indexes, sodass z.B. nur an bestimmten Hochschulen gesucht werden kann oder der Index nach vorher festgelegten Kriterien eingeschränkt werden kann.

Möglich ist auch die externe Speicherung der Daten in einer Datenbank. Dann werden die Daten nur indexiert und durchsuchbar gemacht, jedoch nicht im Index gespeichert. Die Rohdaten werden dann immer aus der Datenbank abgefragt. Dies reduziert die Größe des Index, schränkt aber die Perfomance ein.

3.4. Öffentlichkeit und Datenschutz

Die Suchmaschine soll eigebettet für alle Nutzer zugänglich sein, auch wenn er kein Student ist. Daher dürfen keine sensiblen Daten, wie z.B. E-Mailadressen ausgegeben werden. Diese

Daten müssen also vorher herausgefiltert werden. Zu beachten sind hier auch die Datenschutzbestimmungen der Portale und die damit verbundenen Copyright-Regeln.[xiii] [xiv]

3.5. Nutzerfreundlichkeit/Usability

Der Studieninteressierte soll für den Kurs und die Hochschule begeistert werden. Daher muss das Suchinterface, als auch die Darstellung der Ergebnisse ansprechend sein. Für das Suchinterface kommen eine einfache Suchzeile und ein Suchbutton in Frage, da dies von anderen Websites vertraut ist. Die Darstellung der Ergebnisse sollte auf das notwendige reduziert sein. Durch Auswählen kann der Suchende dann weiter Informationen bekommen und ggf. ähnliche Kurse angezeigt bekommen. Dabei sollte auf eine einfache und intuitive Navigation gesetzt werden. Ein genauer Entwurf der Darstellung ist noch mit dem Project Owner abzustimmen. Im Vordergrund sollte der beworbene Kurs und die anbietende Hochschule stehen.

Für den Administrator sollte eine einfache Art und Weiße bereitgestellt werden die Applikation einzurichten und die Datenquellen zu definieren.

---

[i] https://de.wikipedia.org/wiki/Webserver

[ii] https://www.elastic.co/guide/en/elasticsearch/guide/current/inverted-index.html

[iii] Konchady, M.; Building Search Applications Lucene, LingPipe and Gate; Mustru Publishiong Oakton Virginia; 2008

[iv] https://secure.php.net/manual/de/index.php

[v] http://lucene.apache.org/

[vi] http://framework.zend.com

[vii] http://lucene.apache.org/solr/

[viii] https://de.wikipedia.org/wiki/Representational_State_Transfer

[ix] http://www.rumours.co.nz/manuals/using_moodle_web_services.htm [x] https://demo.bps-system.de/olatce/restapibeta/api/doc

[xi] https://www.danielfett.de/internet-und-opensource,artikel,web-sicherheit

[xii] Konchady, M.; Building Search Applications Lucene, LingPipe and Gate; Mustru Publishiong Oakton Virginia; 2008

[xiii] https://moodle2.uni-leipzig.de/mod/page/view.php?id=5

[xiv] https://bildungsportal.sachsen.de/opal/dmz/1%3A3%3A4009273265%3A1%3A0%3Acid%3Afooter.disclaimer/

https://entwickler.de/online/lucene-ein-suchindex-in-der-praxis-129799.html