

Suchen und Finden in heterogenen Datenbeständen

Technische Informationen zur Textanalyse und Suchfunktion von *Odalis*



Odalis ist ein Framework für webbasierte Portal- und Informationsdienste auf der Basis von *PHP/MySQL*. Wesentlich für *Odalis* ist die Fähigkeit zur semantischen Vernetzung von und intelligenter Suche nach Daten verschiedener Art. Das System indexiert heterogene Datenbestände und bietet dem Nutzer eine komfortable Suchfunktion, mit der Inhalte unterschiedlicher Herkunft im geschützten Netzwerk recherchiert werden können. *Odalis*-Netzwerke können eine beliebige Größe haben (lokal, überregional, weltweit). Suchergebnisse werden als XHTML, XML, serialisiertem PHP-Arrays oder reinem Text ausgegeben: Dies ermöglicht die nahtlose Integration in externe Anwendungen wie etwa Content Management Systeme.

Die hohe Leistungsfähigkeit der von *Odalis*-Systemen beruht auf einer intelligenten Tabellenkonstruktion und einem angereicherten Index: Schon bei der Datenerfassung werden Sprachmuster und semantische Beziehungen analysiert. Auch bei großen Datenbeständen wird so eine hohe Suchgeschwindigkeit von relevanten Ergebnissen erreicht.

Indexierung

Odalis Systeme können zeit- oder ereignisgesteuert unterschiedliche Dateiformate öffnen, deren Inhalte auslesen und indexieren. Bei der Indexierung werden die Volltexte der Dateien in Worte und zusammenhängenden Wortgruppen zerlegt. Anschließend bewertet die Suchmaschine Worthäufigkeit, -position, -länge sowie das semantische Gewicht im Dokument – falls solche Informationen in Form von Formatangaben zur Verfügung stehen (z. B. Auszeichnung von Überschriften). Einzelne Begriffe, Begriffslisten oder Begriffszusammenhänge können zur Gruppierung von neuen Dokumenten dienen, die automatisch innerhalb einer hierarchischen Verzeichnisstruktur einsortiert werden.

Der Thesaurus als Kartografie von Informationsstrukturen

Der Wortindex wird als Thesaurus organisiert und bietet damit Informationen über hierarchische, assoziative oder äquivalente Begriffsbeziehungen, Flexionsformen, Mehrwortbegriffe, Wortfelder usw. Der in der Standardversion mitgelieferte allgemeinsprachliche Thesaurus kann durch Fachthesauri ergänzt oder ausgetauscht werden. Auf diese Weise können Wissensdomänen terminologisch kontrolliert und kartografiert werden (z. B. betriebswirtschaftliches Vokabular). Bei Suchanfragen werden Wortvarianten oder Synonyme des gesuchten Begriffs miteinbezogen und dadurch thematische Zusammengehörigkeiten erkannt. Auch Homonyme (umgangssprachlich: Teekesselchen) können berücksichtigt werden: So lassen sich beispielsweise Dokumente gruppieren, je nachdem, ob der Begriff „Absatz“ im Kontext von „Textformatierung“ oder in betriebswirtschaftlichem Zusammenhang vorkommt.

Zusätzlich unterstützen phonetische Algorithmen die Suche, indem Falschschreibungen erkannt werden: Das System bietet dem Suchenden in diesem Fall ähnliche Begriffe zur Auswahl an (umgangssprachlich: 'Meinten-Sie-Suche').

Sämtliche Bewertungen von Inhalten sind parametrisierbar und können an die jeweiligen Besonderheiten der zu analysierenden Daten und an den Bedarf der Nutzer angeglichen werden.

Matching

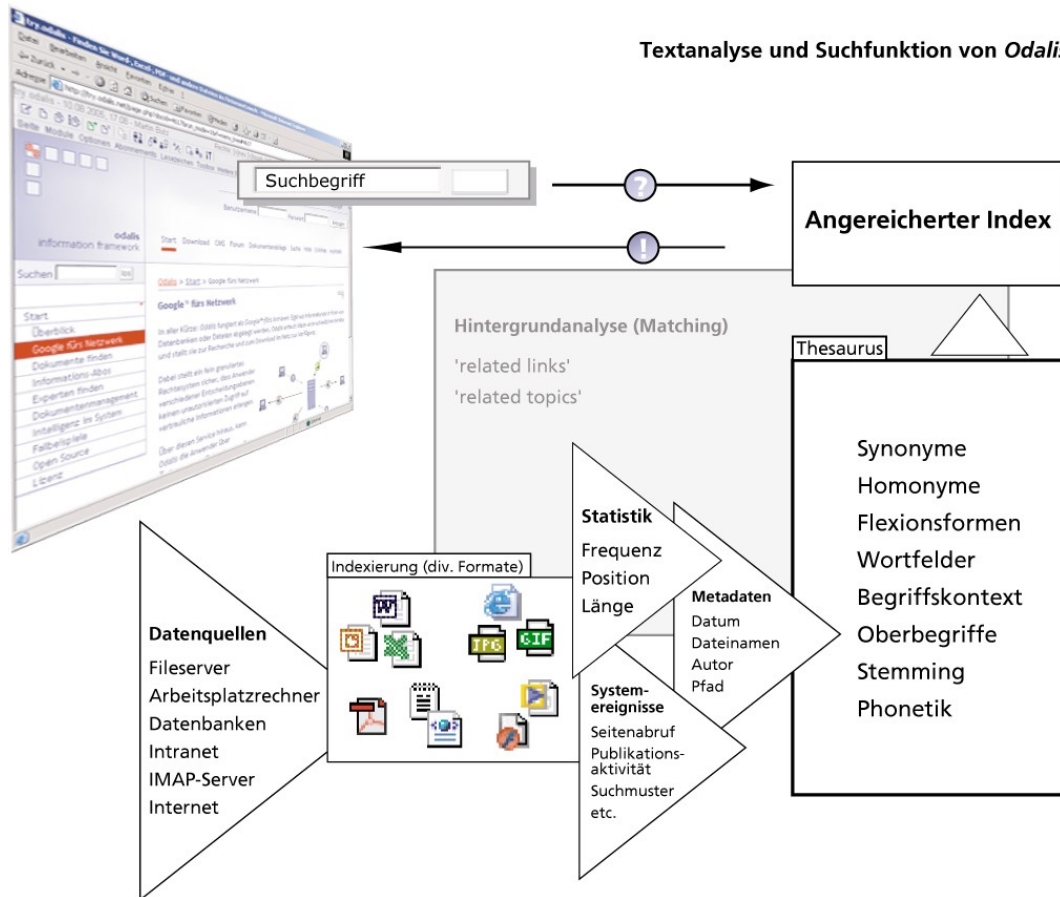
Der Thesaurus unterstützt die Textanalyse und Suche über den gesamten Datenbestand: *Odalis*-eigene oder externe Datenbanken, Internetangebote, Fileserver mit unterschiedlichen Dokumentenarten oder das eigene Intranet. Dies ermöglicht beispielsweise eine produktive Verknüpfung von Expertenprofilen mit dem indexierten Dokumentenbestand: Zusätzlich zu den Dokumenteninhalten kann *Odalis* alle zur Person erfassten Angaben in die Analyse mit einbeziehen (z. B. Wissensgebiete, Kompetenzen, Projekterfahrungen etc.).

Ergänzend werden dynamische Informationen zur Relevanzbewertung von Suchergebnissen herangezogen werden (Ereignisprotokoll des Systems): Hierzu zählen beispielsweise die Publikationsaktivität von Autoren, die Leseaufrufe und die Häufigkeit und Qualität der Verlinkung von Dokumenten.

Als Ergebnis einer solchen Auswertung findet der Suchende nicht nur Hinweise auf Webseiten, Dokumente, Forenbeiträge, Emails und Einträge in Wissensdatenbanken, sondern auch den Kontakt zu Experten. Damit erhält der Nutzer den entscheidenden Vorteil, auch nicht dokumentiertes Wissen persönlich abzufragen.

Sicherheit

Daten werden nur an authentifizierte Benutzer ausgeliefert. Eine Authentifizierung kann über die *Odalis*-eigene Benutzerverwaltung oder extern (z. B. LDAP) erfolgen. Bei Bedarf werden auch alle Login-Informationen verschlüsselt übermittelt (HTTPS/SSL). Bei höchsten Sicherheitsansprüchen wird *jede* Datenauslieferung kryptografisch verschlüsselt (Private-Key-Verfahren). Natürlich kann auch ein *Odalis*-System über ein Virtual Private Network genutzt werden.



Aktuell unterstützte Formate

- PDF
- Text
- RTF
- HTML/XHTML
- Microsoft Word, Excel und PowerPoint
- Alle Medienformate wie Flash, AVI, MPEG, GIF, JPEG etc.

Datenquellen (Odis Repository Server)

- Fileserver (Windows, Unix, Linux)
- Arbeitsplatzrechner (Windows, Unix, Linux)
- Odis-interne Datenbanken
- externe SQL-Datenbanken
- IMAP-Server
- externe, über das Internet erreichbare Informationsangebote (z. B. www.wirtschaftswoche.de)

Kernpunkte der Textanalyse durch die Odis Search Engine

- phonetische Analyse (zur Identifikation fehlerhafter Schreibweisen, 'Meinten-Sie-Suche'; Methode: Eigenentwicklung und *Metaphone*)
- statistisch-linguistische Auswertung von Wortfrequenz, -länge, position (Basisanalyse)
- Stemming (Methode: einfache Wortreduktion, *Porter*)

- Bewertung der Ähnlichkeitsbeziehungen von Texten (Methode: *Levenshtein*)
- Unterstützung durch Thesaurus (hierarchische, assoziative oder äquivalente Begriffsbeziehungen, Flexionsformen, Mehrwortbegriffe, Wortfelder)
- Auswertung von Metadaten (Autor, Dokumentenumfang, Erstellungs- und Änderungsdatum, Schlüsselworte und Bemerkungen z. B. aus MS Worddateien etc.)
- Auswertung von Systemereignissen (Page Views, Verlinkungen, Editiervorgänge etc.)

Suchfeatures

- Volltextsuche mit booleschen Operatoren (UND/ODER)
- Suche nach zusammenhängenden Phrasen
- Unscharfe Suche, 'Meinten-Sie-Suche'
- Sprachspezifische Suche
- Mehrsprachige Suche (Suche nach „Pferd“ findet auch Texte mit „horse“)
- Suche in thematisch gegliederten Teilbereichen (z. B. Unterverzeichnisse einer Dokumentenablage)
- Möglichkeit zum Speichern von Suchanfragen
- Informations-Abonnements für bestimmte Bereiche und/oder mit bestimmten Suchbegriffen (automatische Benachrichtigung bei Änderungen im Dokumentenbestand, vergleichbar den *Google Alerts*)
- Ergebnisansicht sortierbar nach Relevanz, Dokumententyp, Autor etc.
- Baumansicht (Explorer-artig) zur schnellen Orientierung und Verortung der Inhalte
- optional: Wissenslandkarte (Visualisierungen von Sinnfeldern und thematischen Clustern)