INTRAFIND

TopicFinder

Vollautomatische Textklassifikation – zur Themenerkennung und Dokumentenanalyse

KUNDENZITAT

„IntraFind steht für technologisch hochwertige Produkte, Zuverlässigkeit und Kundennähe."

Thomas Müller
Leiter Handelsmarketing, Deutschland
AUDI AG

KEY BENEFITS

- Hohe Zeit- und Kostenersparnis durch Wegfall der manuellen Verschlagwortung
- Amortisation in nur einem Jahr möglich
- Bestes verfügbares, maschinelles Lernverfahren

Inhalte besser verstehen lernen

Die Textklassifikation ist eine zentrale Technologie für die Analyse von Dokumenten und die Identifizierung von Themen. Ein Thema ist als ein semantischer Kontext zu verstehen, welcher allerdings allzu oft nicht oder nicht eindeutig auf der Wortebene zu bestimmen ist. Viele Wörter der deutschen Sprache sind mehrdeutig, d.h. die exakte Bedeutung ist immer abhängig vom jeweiligen Kontext. Dies gilt natürlich auch für andere Sprachen.

In anderen Worten, für die korrekte Interpretation von Informationen sind Wörter zusammen mit dem jeweiligen, lokalen Kontext essentiell. Für das Finden relevanter Information ist die kontextsensitive Interpretation von Inhalten oft entscheidend.

Klassische Enterprise Search- und Information Retrieval-Lösungen stoßen hier an ihre Grenzen. Die Textklassifikation ist eine Schlüsseltechnologie, um jenseits der Wortebene Themen und Kontexte von Dokumenten zu bestimmen und nutzbar zu machen. Zur Bestimmung der Themen werden nicht einzelne Wörter, sondern automatisch generierte Mengen an Wörtern und Mehrwortbegriffen genutzt.

Automatisches Klassifizieren von Inhalten mit dem TopicFinder

Der TopicFinder ordnet bestehende und neue Inhalte automatisch und sehr zuverlässig vorgegebenen Themengebieten (Kategorien) zu.

Diese Klassifikation funktioniert auch auf komplexen, hierarchischen Kategoriensystemen (=Taxonomien) mit einer großen Menge an definierten Kategorien auf verschiedenen Ebenen. Der Einsatz des TopicFinder als automatisches Lernverfahren gliedert sich in zwei Phasen: Trainings- und Produktivphase.

In der Trainingsphase ermittelt der TopicFinder auf Basis von Beispieldokumenten pro Kategorie ein Klassifikationsmodell und erlernt auf diese Weise, Texte inhaltlich zu verstehen. Dies erfolgt unter Verwendung von neuesten Algorithmen und Verfahren des Maschinellen Lernens. Ein Schlüssellernverfahren des TopicFinder ist die Support Vector Maschine (SVM), ein ausgesprochen zuverlässiges und genaues Lernverfahren, wie zahlreiche wissenschaftliche Veröffentlichungen beweisen.

Mittels eines Klassifikationsmodells ist der TopicFinder in der Produktivphase in der Lage, neue Dokumente oder Inhalte automatisch den definierten Kategorien zuzuordnen. Je nach konkreter Anwendung können Dokumente aufgrund dieser Zuordnungen spezifisch selektiert und eingeordnet sowie automatisch weitergeleitet oder annotiert werden.

Vorteile des TopicFinder

- Massive Zeit- und Kostenersparnis durch Wegfall der manuellen Verschlagwortung und Themenzuordnung.
- Amortisation der Lösung innerhalb des ersten Jahres!
- Bestes verfügbares, maschinelles Lernverfahren.
- Qualitativ hochwertigere und zuverlässigere Ergebnisse als rein statistisch arbeitende Klassifikationswerkzeuge aufgrund der Kombination von linguistischer Vorverarbeitung und neuesten mathematischen Algorithmen.

ANWENDUNGSBEISPIELE

- **Themenbasierte Suche** für die personalisierte Informationsbereitstellung – ein benutzerspezifischer TopicFinder filtert nur die Informationen heraus, die dem individuellen, inhaltlichen Profil des Benutzers entsprechen.
- **Automatische Verschlagwortung** von neuen Informationen oder Nachrichtenartikeln in Verlagen und Nachrichtenagenturen, z.B. gemäß der IPTCKategorisierung. Anstelle von langwieriger manueller Verschlagwortung und Indexierung durch Fachpersonal.
- **Automatische Zuordnung** von Produkten zu Produktkategorien, z.B. für die standardisierten Klassifikationssysteme eClass und UN/SPSC auf Basis der Hersteller - Produktdatenblätter.
- **Klassifikation eingehender digitaler Post,** Verteilung von EMails in einem Support-Center anhand der inhaltlichen Analyse und dem Matching mit dem passendsten Sachbearbeiter.
- **Abbildung des Posteingang-Workflows,** angefangen von dem Scannen und der OCRAufbereitung von Poststücken, gefolgt von der automatischen Klassifikation des Posteingangs in Dokumentklassen (z.B. Auftragsbestätigung, Rechnung, Vertrag, Memo, Bericht, Gutachten) mit anschließender Verteilung an den Sachbearbeiter. Textsprache und Textinhalt dienen der Steuerung des digitalen Dokumentenflusses. Mit Hilfe unseres Produktes NAMER können dann in einem weiteren Schritt Kerninformationen (z.B. Name, Adresse, Vertrags- oder Kundennummer) sicher aus dem Dokument extrahiert werden.
- **Automatische inhaltliche Erstellung eines Newsletters:** Agenturmeldungen bzw. gekaufte Fachartikel werden automatisch entsprechend eines für den jeweiligen Newsletter erzeugten Profils gefiltert und einzelnen Kapiteln des Newsletters zugeteilt. Der TopicFinder analysiert diese Artikel, sortiert sie nach Signifikanz und gruppiert ähnliche Artikel oder gar Dubletten zusammen, um die Weiterverarbeitung zu vereinfachen und die Informationsflut übersichtlich zu gestalten.

Produkteigenschaften

Umfangsreiche Trainings- und Administrations- GUI:

- Datenanalyse, Training, Evaluierung und Testen direkt über den Web Browser mittels Web- Anwendung
- Gleichzeitiges Arbeiten mehrerer User über die Web-Anwendung
- Effiziente Analyse der Trainingsdaten- und Testdatenmenge hinsichtlich von Konflikten und Inkonsistenzen der Zuordnung (z.B. Identifizierung sehr ähnlicher Dokumente mit unterschiedlicher Themenzuordnung)
- Training und Evaluierung von Klassifikationsmodellen
- Aufbau und Wartung von komplexen, hierarchischen Taxonomien
- Gleichzeitige Verwaltung von mehreren Kategoriensystemen und Trainingsdatenmengen

Mehrstufige Vorverarbeitung der Dokumente:

- Unterstützte Dokumenttypen: ASCII, MS Office, PDF, HTML, XML
- Linguistische Vorverarbeitung für Texte in den Sprachen DE, EN, FR, ES, IT, u.a. durch Lemmatisierung (Grundformzeugung) & Kompositazerlegung, Tokenizer für Chinesisch
- Auffinden inkonsistenter Trainings- und Testdaten beim Dokumentenimport
- Texte aus OCR-Quellen werden zur Qualitätsverbesserung n-gram-vorverarbeitet
- Ausgereifte statistische und informationstheoretische Verfahren zur Merkmalsselektion

Training von Klassifikationsmodellen mit zahlreichen Wahlmöglichkeiten:

- Einsatz der Support Vector Maschine (SVM) sowie weiterer Lernverfahren
- Optimierungsoptionen während des Trainings zur automatischen Bestimmung von optimalen Lernparametern
- Vermeidet schlechte Generalisierung (das zentrale Problem anderer Verfahren ist die Überspezialisierung auf die Trainingsdaten und damit schlechte Generalisierung)
- Zuordnung eines Dokuments zu mehr als einer Kategorie möglich, optional kann aber auch eine eindeutige Klassifikation erzwungen werden.
- Hohe Skalierbarkeit des Trainings und der Cross-Validierung durch den Einsatz von Multi- Threading

Umfangreiche Test- und Evaluierungsmöglichkeiten zur Ermittlung der Qualität des Klassifikationsmodells:

- Cross-Validierung für die Bestimmung der Qualität eines Klassifikationsmodells
- Effiziente Analyse der Zuordnung mittels True/ False Positives/Negatives
- Dokumenten-Clustering

In der Produktivphase:

- Effiziente Realisierung der Klassifikation im Batch-Modus
- Einbindung von Thesauri (Synonymrelationen) zur Klassifikation
- Verbesserung der Klassifikation durch Kombination mit fallbasierter Klassifikation auf den Trainingsdaten
- Klassifikation von Dokumenten in unterschiedlichen Sprachen möglich

Weitere Produkteigenschaften:

- Komplett in Java implementiert (plattformunabhängig), umfangreiche Java-API verfügbar
- Basierend auf der Open Source-Volltextsuchtechnologie Apache Lucene
- Editierfunktion für erlernte Klassifikationsmodelle
- Eigene Verwaltung von Benutzerfeedback für Verbesserung der Klassifikationsqualität

SYSTEMVORAUSSETZUNGEN

Das Trainieren von Klassifikationsmodellen ist abhängig von der Komplexität der Taxonomie und der Anzahl der Trainingsdokumente eine rechenintensive Aufgabe, die entsprechende Hardware voraussetzt.

Durch die Aufteilung von rechenintensiven Aufgaben auf mehrere Threads wird die Skalierbarkeit des TopicFinder deutlich erhöht und es können mehrere hunderttausend Dokumente sowie komplexe Taxonomien trainiert werden.

Empfohlen wird grundsätzlich ein Server mit zwei Quadcore-Prozessoren sowie mindestens 8 GB Hauptspeicher. Je nach Komplexität der Taxonomie oder der Größe der Trainingsdatenmenge kann dies aber auch abweichen. Der TopicFinder ist eine reine Java- Anwendung und benötigt neben einer obligatorischen Java JRE auch einen Servlet Container wie bspw. Apache Tomcat.

Angebotsspektrum

- Lizenzierungen (Lizenzkauf oder Miete, volumenabhängige Abrechnung ohne Fixkosten)
- OEM-Programm
- Aufbereitung der Dokumente für die Trainingsphase, Durchführung des Lerntrainings
- Consulting und Schulung für Ihre Projekte
- Kundenspezifische Entwicklungen und Integration unserer Produkte in Ihre IT-Infrastruktur
- Softwarepflege

UNTERNEHMENSPROFIL

- Die IntraFind Software AG, gegründet im Jahr 2000, ist ein unabhängiger Softwarehersteller mit Firmensitz in München.
- IntraFind entwickelt Produkte und Lösungen für das effiziente Suchen, Finden, Analysieren von strukturierten und unstrukturierten Informationen unter Berücksichtigung aller verfügbaren Datenquellen eines Unternehmens.
- Volltextsuche und die komplette Bandbreite an Verfahren für Textanalyse, Machine Learning, Deep Learning, Natural Language Processing und Artificial Intelligence, kombiniert mit den Möglichkeiten von Graphdatenbanken für Big Data Analytics, bilden hierbei den Schwerpunkt.
- Namhafte Kunden von IntraFind sind: AUDI AG, BMW AG, Bundeswehr, IHK Berlin, MAN Truck & Bus AG, MTU Aero Engines AG, Robert Bosch GmbH, Rohde & Schwarz GmbH & Co. KG, ZF Friedrichshafen AG.

IntraFind Software AG
Landsberger Straße 368
80687 München
Tel: +49 (89) 309 0446-0
www.intrafind.de

IntraFind entwickelt Produkte für das effiziente Suchen, Finden, Analysieren von Informationen unter Berücksichtigung aller Datenquellen eines Unternehmens.