Diversität im Web

MEINUNGSVIELFALT SICHTBAR UND NUTZBAR MACHEN

Im Internet gibt es unzählige Webseiten mit Texten aller Art. Das Projekt LivingKnowledge beschäftigt sich mit dieser Vielfalt an Informationen und hat zum Ziel, eine neue Generation von Such- und Wissensmanagementtechnologien zu entwickeln, die ihre Resultate dem Nutzer einheitlicher, vollständiger und in ihrem Kontext darstellen.

1

Wissen und dessen Verbreitung werden stark beeinflusst von der Vielfalt (Diversität) kultureller Voraussetzungen, unterschiedlichster Ideologien sowie des zeitlichen Zusammenhangs. Themen wie Immigration, Treibhauseffekt oder die Präsidentschaftswahl in den USA 2008 werden im Web durch eine Vielzahl unterschiedlicher Quellen behandelt. Neben eher traditionellen Quellen wie Portale großer Nachrichtenanbieter wie BBC oder CNN bieten Blogs sowie soziale Web 2.0 Systeme wie Twitter, Flickr oder YouTube nun einer großen Menge von Webbenutzern die Möglichkeit, der Web Community eigene Informationen und Meinungen auf unkomplizierte Weise bereitzustellen. Weiterhin schaffen diese Webumgebungen die Möglichkeit zu intensiver Interaktion zwischen Benutzern in Form von Diskussionen, durch Kontaktaufnahme zu anderen Benutzern oder die Bildung von Gruppen. Mit dieser sozialen Interaktion zwischen Benutzern im Web beschäftigt sich das junge Forschungsgebiet »Web Science«.

Auf der einen Seite bietet die Vielfalt von unterschiedlichen Quellen die Möglichkeit, aus verschiedenen Blickwinkeln Informationen über Ereignisse zeitnaher zu erhalten. Auf der anderen Seite jedoch werden User mit einer Vielzahl von divergierenden Ansichten sowie Widersprüchen in dargestellten Sachverhalten und Informationen konfrontiert.

Bei polarisierenden Themen wie Immigration spielen unterschiedliche Sichtweisen (zum Beispiel rechtspopulistisch gegenüber liberal) eine starke Rolle, was sich in kontroversen Diskussionen und oft widersprüchlichen Darstellungen und Interpretationen von Statistiken (zum Beispiel Kriminalitätsraten) oder (vermeintlichen) Fakten widerspiegelt. Um diese stark anwachsenden Informationsmengen im Web verwenden und analysieren zu können, ist die Erarbeitung von Techniken zur zuverlässigen Auswertung von Informationen aus unterschiedlichen Quellen unabdingbar.

Im Bereich des Text Retrievals wurde Diversität bisher nur unter dem Problem der Diversifizierung von Suchergebnissen betrachtet. Bestehende Ansätze kombinieren dazu Maße der Diversität und der Ähnlichkeit, um die Qualität der Suchergebnisse zu verbessern, das heißt, einerseits möglichst viele relevante Ergebnisse bereitzustellen, andererseits aber auch eine große Vielfältigkeit anzubieten.

LivingKnowledge

Die Vision des aktuell am Forschungszentrum L3S laufenden EU FET Projektes LivingKnowledge ist es, Diversität als Bereicherung zu betrachten und diese nachvollziehbar, verständlich und nutzbar zu machen. Obgleich moderne Suchmaschinen ein beachtliches Maß an Effizienz und Effektivität vorweisen können, werden deren Benutzer häufig mit langen, linearen Listen von Suchresultaten konfrontiert. Die Suchresultate werden zwar unter Verwendung von recht fortschrittlichen Algorithmen sortiert, Ziel dieser Algorithmen ist jedoch häufig nur das Extrahieren der populärsten Suchergebnisse. Zudem erhält man selbst für Sachfragen wie »Wie groß ist die Bevölkerungszahl der USA?« oder »Wie viele neue Kraftwerke werden im nächsten Jahr in Deutschland ans Netz angeschlossen?« Dokumente als Suchresultate, die oft nur implizite und inkonsistente Antworten enthalten. Somit scheitert die aktuell vorhandene Such- und Wissensverarbeitungstechnologie häufig an den Bedürfnissen der Benutzer, zudem wird die Diversität von Informationen und deren Entwicklung nicht berücksichtigt. Die Herausforderungen bestehen nun darin, Informationen aus unterschiedlichen Quellen im Web zu suchen, zu vereinheitlichen und dem Benutzer einen integrierten und umfassenden Überblick zu bieten. Hinzu kommt die selbst für erfahrene Nutzer schwierige Aufgabe, Informationen auf ihre Vertrauenswürdigkeit und Vollständigkeit hin zu überprüfen, da grundsätzlich die Gefahr besteht, falsch oder

Abbildung 1
Diversität und Vielfalt spielen im Web eine große Rolle.

Abbildung 2
Diversität im Web sichtbar machen.

unvollständig informiert zu werden. Auch der zeitliche Aspekt, der eine andere zentrale Dimension von Diversität darstellt, spielt eine wichtige Rolle. Aktuelle Webverzeichnisse bieten lediglich Schnappschüsse des Webs entlang dieser Dimension, jedoch keine Methoden für die Analyse und das Verständnis von Informationen in ihrem zeitlichen Zusammenhang.

Forschungsfragen

Ziel der Forschung im Bereich Diversität von Informationen und Informationsbeständen am L3S ist es, eine neue Generation von Such- und Wissensmanagementtechnologie zu entwickeln, die Suchresultate konziser, vollständiger und in ihrem Kontext darstellt. Für welche Themen oder Suchanfragen treten welche Aspekte von Diversität vorwiegend auf? Wie gut und repräsentativ werden diese Aspekte von den führenden Suchmaschinen abgedeckt? Wie ändern sich Meinungen zu bestimmten Themen mit Zeit und Ort? Zu welchen Themen findet man widersprüchliche Aussagen? Wie lässt sich die Zuverlässigkeit unterschiedlicher Informationsquellen beurteilen? In welcher Form präsentiert man dem Benutzer Inhalte und Suchergebnisse mit hoher Diversität?

Dies sind einige der Forschungsfragen, denen wir am L3S nachgehen. Die Forschung und die interdisziplinäre Kombination von Resultaten umfasst die Bereiche der Informationsextraktion, zeitliche Evolution von Wissen, Diversität und Polarisation von Information, Clustering und Aggregation von Wissen sowie erweiterte Suchtechnologie. Informationsextraktion beinhaltet die Extraktion von Fakten und Meinungen aus Textdaten sowie aus einer Kombination von Texten und Bildern. Wissensevolution befasst sich mit der zeitlichen Entwicklung von Wissen in all seinen Facetten, seien es Fakten, Sprache, Meinungen oder Wissensstrukturen wie Ontologien. Der Bereich der Diversität und Polarisation befasst sich mit dem Erkennen und Analysieren von unterschiedlichen Meinungen, widersprüchlichen Informationen und polarisierenden Themen. Clustering und Aggregationstechniken werden angewendet, um dem Benutzer eine strukturierte und konzise Übersicht zum Decken seines Informationsbedarfs zu liefern. Aufbauend auf den genannten Teilbereichen schließlich, bieten moderne und erweiterte Suchtechnologie innovative Lösungen für das Auffinden von Informationen sowie Navigieren und Browsing von großen Datenmengen, die dabei die Diversität und temporale Aspekte von Daten und Wissen mit einbeziehen, und somit dem Endbenutzer umfassendere und zuverlässigere Informationen liefern.

Abbildung 3
Meinungsvielfalt als Chance verstehen.

Abbildung 4
Zeitliche Entwicklung der öffentlichen Meinung für die Präsidentschaftskanditaten Obama und McCain 2008 (Schätzungen nur aufgrund der Blogosphere im Vergleich zu traditionellen Umfragewerten).

Ausgewählte Ergebnisse

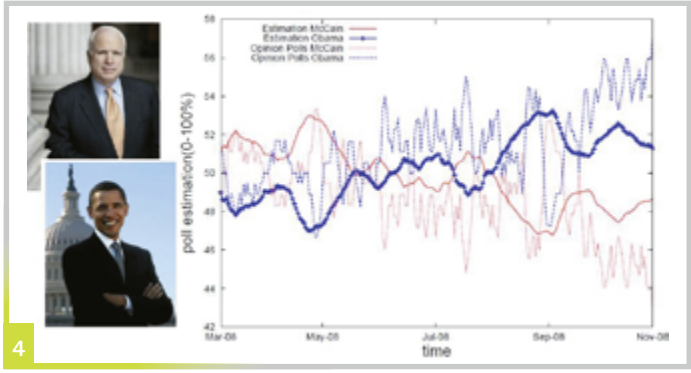
Im Laufe des Projekts hat das L3S zusammen mit seinen Partnern verschiedene Aspekte von Diversität im Web analysiert und Technologien zur Nutzbarmachung entwickelt. Die Analyse von Kommentaren im Web 2.0 und dem Social Web zusammen mit benutzergenerierten Benotungen und Bewertungen stellen einen wichtigen Beitrag zur verbesserten Informationsbereitstellung für die einzelnen Nutzer dar. Implizites Wissen über Inhalte, Klassen und Gruppierungen, oder gemeinschaftliche Interessen können effektiv gesammelt und ausgewertet werden. Eine konkrete Fragestellung in diesem Zusammenhang ist beispielsweise, wie nützlich die Kommentare der Benutzer auf der Video-sharing Plattform YouTube sind, um Vorhersagen treffen zu können betreffs der Akzeptanz von zukünftigen Kommentaren und Nutzerrückmeldungen. Methoden des maschinellen Lernens helfen, Modelle für die Kommentare der Nutzer zu entwickeln. So können in Anwendungen beispielsweise die interessantesten Kommentare von Benutzern zu einem Video eingeblendet werden.

Ein weiteres Beispiel unsere Forschung in diesem Bereich

ist das Analysieren von Nachrichtentexten. Nachrichtenseiten im Web werden immer beliebter. Die Möglichkeit der zeitnahen Bereitstellung und Interaktion mit dem Leser machen Onlinenews sowohl für Anbieter als auch Leser interessant. Die Vielfalt und Menge an Informationen, die dadurch weltweit tagtäglich entstehen, machen das Vorfiltern und Personalisieren von Daten für den individuellen Leser notwendig.

Ein Forscherteam am L3S hat sich mit dieser Problemstellung beschäftigt und Lösungen erarbeitet, welche auf das Extrahieren von latenten Faktoren in Nachrichtentexten abzielen. Diese latenten Faktoren, wie beispielsweise das Erwähnen bestimmter Themen oder Personengruppen in einem Text, können auf wichtige oder unwichtige Nachrichten hindeuten. Die Nennung spezifischer Entitäten, wie zum Beispiel »Präsident Obama«, kann ein Hinweis auf einen wichtigen Nachrichtenartikel sein, wogegen das Vorhandensein verschiedener anderer Wörter, zum Beispiel »Lokalpolitik« auf global unwichtige Nachrichten schließen lässt. Eine breite Analyse der entdeckten Faktoren gibt weiterhin Aufschluss über globale Ereignisse und erlaubt im Zusammenhang mit der zeitlichen Betrachtung eine Klassifizierung von Events sowie Vorhersagen über die Wichtigkeit zukünftiger Artikel. Maschinelles Lernen wird eingesetzt, um aus vergangenen Zeitungsartikeln ein Modell zu entwickeln, welches die Wichtigkeit zukünftiger Nachrichten abschätzen kann. Das Modell kann so lernen, dass ein Artikel, der die Wörter »Krieg«, »viele Tote«, und »UNO« in Kombination beinhaltet, wichtigere Informationen enthält als ein Text mit den Wörtern »mögliche Steuererhöhung«, »Diskussion« und »Uneinigkeit«.

Neben professionellen Nachrichtenartikeln gibt es im Internet auch Blogs, von Internetnutzern geführte online Tagebücher. Diese Texte enthalten oft subjektive Informationen, Meinungen, und Erfahrungen. Innerhalb des LivingKnowledge Projektes haben wir uns mit Blogs und Sentimentanalyse im Zusammenhang mit der Präsidentschaftswahl in den USA 2008 auseinandergesetzt. Wir konnten zeigen, dass das automatische Extrahieren von positiven und negativen Meinungen der beiden Kandidaten eine gute Prognose erlaubt. Mit vergleichsweise geringem Aufwand ermöglicht die Analyse von Blogs eine schnellere und qualitative gleichwertige Auswertung der öffentlichen Meinung eines Kandidaten im Vergleich zu traditionellen Meinungsumfragen vor einer Wahl.



4

Dr. Ralf Krestel
Jahrgang 1980, Studium der Informatik und Philosophie in Karlsruhe und Montreal. Anschließend Promotion an der Leibniz Universität Hannover. Arbeitete von 2007 bis 2012 am L3S und ist zurzeit wissenschaftlicher Mitarbeiter an der Universität von Kalifornien in Irvine. Seine Forschungsschwerpunkte liegen in Textmining, Natural Language Processing und Information Retrieval. Kontakt: krestel@l3s.de

Dr. Stefan Siersdorfer
Jahrgang 1976, promovierte 2005 in der Arbeitsgruppe für Datenbanken und Informationssysteme am Max-Planck-Institut für Informatik. Anschließend war er Research Fellow an der University of Sheffield und Gastwissenschaftler bei Yahoo! Barcelona. Momentan ist er Senior Researcher und Projektleiter am Forschungszentrum L3S. Seine Forschungsinteressen umfassen Verteiltes Machine Learning sowie Suche und Datamining im Social Web, sowie Crawling und Analyse von Web Daten. Kontakt: siersdorfer@l3s.de