Kollaboratives Strukturieren von Multimediadaten für Peer-to-Peer Netze

1. **PG-Thema:** Kollaboratives Strukturieren von Multimediadaten für Peer-to-Peer Netze
2. **PG-Zeitraum:** WS 2004/2005 und SS 2005
3. **PG-Umfang:** jeweils 8 SWS
4. **PG-Veranstalter:**

   - Katharina Morik, Lehrstuhl Informatik VIII, morik@ls8.cs.uni-dortmund.de, GB IV/R115, Tel. 5101
   - Michael Wurst, Lehrstuhl Informatik VIII, wurst@ls8.cs.uni-dortmund.de, GB IV/R120, Tel. 6490

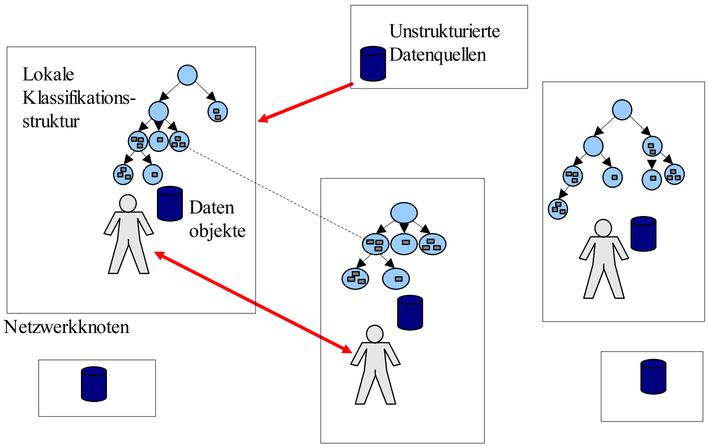
5. **PG-Aufgabe:**

   Zusammenfassung

   Im Rahmen der Projektgruppe soll ein System zur dezentralen Verwaltung von multimedialen Daten auf Grundlage von peer-to-peer Netzen[1] entwickelt werden. Die Grundidee besteht darin, dass Nutzer ihre Daten lokal und unabhängig von anderen Nutzern speichern und nach ihrer individuellen Sicht strukturieren/klassifizieren können (Audiodaten z.B. nach Musikgenres). Dabei sollen sie von intelligenten Agenten unterstützt werden, welche wiederum auf die Daten und persönlichen Strukturen anderer Nutzer zugreifen. Auf diese Weise können die Benutzer kollaborativ und verteilt Daten verwalten. Als Basis soll dabei zum einen das Paradigma der peer-to-peer Netzwerke dienen, welches, berühmt durch Anwendungen wie Gnutella oder Freenet, mittlerweile auch in viele Groupwareszenarien (z.B. [10]) Verwendung findet. Zum anderen sollen Methoden des Maschinellen Lernens[3] verwendet werden, die es ermöglichen, große Mengen von multimedialen Datenobjekten effizient zu verarbeiten, d.h. zu klassifizieren oder zu gruppieren. Die Studenten können zur Realisierung des Systems zum einen auf bestehende peer-to-peer Lösungen, wie z.B. JXTA[9] von Sun aufbauen, zum anderen auf Methoden und Algorithmen welche im Rahmen des Yale Systems am Lehrstuhl für Künstliche Intelligenz entwickelt wurden ([8],[7]). Mindestziel der PG ist es, ein System zu entwickeln in dem Benutzer Klassifikationsschemata erstellen und austauschen können. Außerdem soll es möglich sein, automatisch Daten in ein Schema einzuordnen.

   PG-Aufgabe

   Viele aktuelle Systeme zur Verwaltung von Informationen basieren auf einer zentralen Architektur. Dies betrifft zum einen das Bereitstellen der Daten auf einem zentralen Server zum anderen das Strukturieren der Daten durch ein gemeinsames globales Klassifikationsschema. Ein Beispiel für ein solches Klassifikationsschema ist z.B. der Amazon Produktkatalog. Solche globalen Lösungen sind nicht immer zufriedenstellend. Zum einen haben verschiedene Benutzer oder Benutzergruppen häufig verschiedene Vorstellungen davon, wie Daten "richtig" strukturiert werden sollen (z.B. wie Musikgenres in Beziehung stehen und welches Musikstück zu welchem Genre gehört). Zum anderen ändern sich sowohl die Daten als auch die Klassifikationsschemata in vielen Szenarien eher lokal als global. Indem beispielsweise einzelne Nutzer lokal neue Datenobjekte



hinzufügen, wird auch eine lokale Verfeinerung der Klassifikationsschemata nötig, beispielsweise durch das Hinzufügen von Unterkategorien. Diese Dynamik ist durch zentralisierte Systeme nur schwierig einzufangen ([6],[5]). Beide Probleme legen es nahe, die Datenverwaltung dezentral zu organisieren. Während verschiedene peer-to-peer file sharing Systeme (wie Gnutella oder Freenet) zu einer völlig neuartigen und höchst populären Form des dezentralen Anbietens und Suchens von Daten geführt haben, soll an dieser Stelle ein Schritt weiter gegangen werden. Aufbauend auf existierenden peer-to-peer Systemen soll eine dezentrale kollaborative Strukturierung von Daten ermöglicht werden. Die Idee dabei ist folgende. An jedem Knoten des Netzwerks können Datenobjekte gespeichert und in einer individuellen Taxonomie (d.h. einem hierarchischen Klassifikationsschema) strukturiert werden. Jeder Benutzer kann somit eine völlig individuelle Sicht auf die Daten haben, unabhängig von allen anderen Benutzern. Um den Benutzern die Aufgabe der Datenverwaltung zu erleichtern, werden sie bei der Erstellung und Verwaltung ihrer persönlichen Taxonomie von intelligenten Agenten unterstützt. Basis für diese Unterstützung bilden wiederum persönliche Taxonomien anderer Nutzer[2]. Ein typisches Beispiel wäre, dass ein Nutzer die Datenobjekte eines anderen Nutzers mit seiner eigenen Taxonomie strukturieren möchte, um z.B. festzustellen, ob dort für ihn interessante Daten liegen. Es wäre aber auch möglich, dass ein Benutzer eine große Anzahl von bislang unstrukturierten Daten (z.B. Musikstücke, welche in einem Internetradio mitgeschnitten wurden) strukturieren möchte und dazu die Taxonomie eines anderen Benutzers anwendet. Aber auch klassische Aufgaben wie das Auffinden ähnlicher Datenobjekte oder das Charakterisieren von Objekten kann auf diese Weise durchgeführt werden. So können beispielsweise Objekte die häufig gemeinsam in persönlichen Taxonomien vorkommen als ähnlich angesehen werden. Entsprechend können die Bezeichnungen von Kategorien in einer persönlichen Taxonomie als Basis für das Gewinnen von Stichwörtern genutzt werden. Bezogen auf Musik könnten z.B. drei Nutzer ein Musikstück unter "Jazz" subsumieren und ein Nutzer unter "Rock". Da das System die Nutzer nicht zwingt sich auf eine Bezeichnung zu einigen, können diese Meinungen koexistieren und werden nur durch statistische Verfahren vermittelt. "Jazz" würde z.B. in diesem Fall als relevantes Stichwort für das Musikstück ausgegeben, "Rock" hingegen als weniger relevantes.

Das Ziel der Projektgruppe ist entsprechend folgendes: Basierend auf einer existierenden peer-to-peer Plattform, welche Basisdienste wie Dateitransfer und Suche bereitstellt, sollen ein Protokoll und prototypische Agenten erstellt werden, welche kollaborativ und dezentral Nutzern helfen, Daten in hierarchischen Strukturen zu verwalten, sowie neue relevante Datenobjekte zu finden und Beziehungen zwischen Datenobjekten aufzudecken.

Im einzelnen sind folgende Funktionen denkbar:

(a) Klassifikation von neuen Datenobjekten in eine bestehende persönliche Taxonomie
(b) Automatisches Erstellen einer personalisierten Taxonomie aus bislang unstrukturierten Daten
(c) Charakterisierung von Datenmengen durch Begriffe (Stichwörter)
(d) Auffinden interessanter Nutzer, bzw. Knoten im Netzwerk
(e) Empfehlen von Datenobjekten und Begriffen sowie das Auffinden von Beziehungen zwischen Objekten und zwischen Begriffen

Das System soll generisch sein und mit verschiedenen Datentypen umgehen können, allerdings wird der Schwerpunkt auf Multimediadaten liegen. Zur Lösung dieser Aufgabe können die Studenten auf das Yale System[8] zurückgreifen, welches am Lehrstuhl für Künstliche Intelligenz entwickelt wurde. Yale ist eine Umgebung zur Durchführung von Experimenten für Maschinelles Lernen. Neben zahlreichen Klassifikations- und Clusteringverfahren beinhaltet Yale auch Operatoren zur Gewinnung von Merkmalen aus Text und Audiodaten, sowie verschiedene Verfahren zur Auswahl und zur Gewichtung von Merkmalen. Das System ist vollständig Java und XML basiert und kann ohne großen Aufwand in neue Anwendungen eingebunden werden. Desweiteren soll auf ein bestehendes peer-to-peer Netzwerk aufgebaut werden, welches bereits Basisfunktionalität wie Datenaustausch und Suche zur Verfügung stellt. Dieses System soll um die oben genannten Dienste erweitert werden.

Im einzelnen ist folgendes Vorgehen geplant:

(a) Es muss ein peer-to-peer file sharing System ausgewählt werden, welches sich im obigen Sinne erweitern läßt. Besonders bietet sich die Java basierte Plattform JXTA (von Sun) für diese Aufgabe an, da diese bereits eine komplette Opensource Implementierung für das Austauschen und Suchen von Daten, sowie für das dynamische Verwalten von Benutzergruppen bereitstellt.
(b) Aufbauend auf dem gewählten System sollen XML-basierte Datenstrukturen und Kommunikationsprotokolle entwickelt werden, welche den Agenten den Austausch von relevanten Informationen ermöglichen.
(c) Um Verfahren des Maschinellen Lernens anwenden zu können, müssen die Datenobjekte zunächst geeignet kodiert werden. Dazu sind Verfahren der Merkmalsextraktion, welche im Rahmen des Yale Systems am Lehrstuhl für Künstliche Intelligenz entwickelt wurden, anzupassen und einzubinden.
(d) Lernverfahren aus dem Yale System sollen angepasst werden um (eine Teilmenge der) Assistenzaufgaben zu lösen. Diese Verfahren sollen dann in intelligenten Agenten gekapselt werden, welche an jedem Knoten des Netzwerks ausgeführt werden.
(e) Um alle Systemkomponenten zusammenzuführen, soll eine prototypische Anwendungen implementiert werden.

(f) Das fertige System soll intensiv evaluiert werden, vor allem hinsichtlich der Frage, ob ein kollaborativer Ansatz (also das Verwenden von benutzererstellten Informationen) beim Strukturieren von Daten einem rein datenbasierten Herangehen überlegen ist.

(g) Sowohl die Gesamtarchitektur, als auch die einzelnen Komponenten sollen nach Standards der Softwareentwicklung dokumentiert und einer Qualitätssicherung durch die Studenten unterzogen werden. Auf diese Weise wird ein Einblick in professionelle Entwicklung und Verifikation von komplexen verteilten Anwendungen möglich.

6. **PG-Teilnahmevoraussetzungen**

Da für die gegebene Aufgabe verschiedene Kompetenzen nötig sind, sollte möglichst *eine* der folgenden Veranstaltungen besucht worden sein:

- Künstliche Intelligenz, Maschinelles Lernen, Data Mining
- Rechnernetze und verteilte Systeme

7. **Minimalziel**

Das Minimalziel ist die Erstellung eines Systems, welches das lokale Speichern und Verwalten von Daten in Form von persönlichen Taxonomien erlaubt, sowie den Austausch von Daten und persönlichen Taxonomien zwischen Nutzern. Außerdem soll mindestens die automatische Klassifikation und das Gruppieren von Datenobjekten möglich sein.

8. Literatur

[1] Milojicic, M.J. et. al: Peer-to-peer computing. Technical Report HPL-2002-57, HP Lab, 2002.

[2] Maes P.:Agents that reduce work and information overload, Communications of the ACM, 1994.

[3] Mitchell, T: Machine Learning, McGraw Hill, 1997.

[4] Anderberg, M.: Cluster Analysis for applications, Academic Press 1973.

[5] Novak, J.; Wurst, M.: Discovering, Visualizing and Sharing Knowledge through personalized learning knowledge maps, in: Agent Mediated Knowledge Management, Springer 2003.

[6] Abecker A.; Bernardi A.; van Elst, L.: Agent Technology for Distributed Organizational Memories: The Frodo Project. ICEIS (2) 2003: 3-10.

[7] Mierswa, I.: Automatisierte Merkmalsextraktion aus Audiodaten. Master's thesis, Fachbereich Informatik, Universitaet Dortmund,2003.

[8] Mierswa, I.; Klinkberg, R.; Fischer, S.; Ritthoff, O.: A Flexible Platform for Knowledge Discovery Experiments: YALE – Yet Another Learning Environment. In LLWA 03 - Tagungsband der GI-Workshop-Woche Lernen - Lehren - Wissen - Adaptivität,2003.

[9] http://www.jxta.org

[10] http://www.groove.net