arbido

2008/4 Informationswissenschaft: Die Instrumente der Zukunft

Vandergheynst Pierre,

Tags: Big Data Elektronisch, Informatik,

Datenkompression und Archivierung: das Binom der Zukunft

Radiowellen, Fernsprechleitungen und Computerkabel transportieren jeden Tag astronomische Mengen von digitalen Daten. Wie soll man diese Daten referenzieren, wenn u.a. Profis der Informationsdokumentation diese Daten archivieren und «gewöhnliche» Nutzer sie nutzen wollen? Die Forschungsgemeinde erwartet eine doppelte Herausforderung: einerseits die Kompression, andererseits die Indexierung.

Fakten

Wenn die Rede ist von «astronomi schen Datenmengen», die tagtäglich mit verschiedenen Geräten und Medien (TV, Telefon, Internet, Überwachungskameras etc.) transportiert werden», so ist das keine Übertreibung – die (nahe!) Zukunft wird uns zeigen, dass «astronomisch» noch um den Faktor n zunehmen wird. Das von der IDC im März 2007 herausgegebene Weiss buch (1) hält fest, dass die Gesamt summe von digitalen Daten, die 2006 produziert wurde, 1,288 x 10 hoch 18 Bytes beträgt, das sind 161 Exabytes oder 161 Billionen Gigabytes; oder mit anderen Worten drei Millionen Mal die Information, die in sämtlichen je geschriebenen Büchern enthalten ist. Das Beste kommt aber noch: Gemäss dem Bericht der IDC wird diese Informationsmasse bis zum Jahr 2010 noch um den Faktor 6 anwachsen.

Damit drängt sich folgende Frage auf: Man weiss, dass 95% der Daten nicht strukturiert sind – wie soll man sie also referenzieren? Die Antwort auf diese Frage ist für die Berufsleute aus dem Bereich Informationsdokumentation von entscheidender Wichtigkeit: Sie werden mit unter den Ersten sein, welche die von den Forschern gegenwärtig zu diesem Zweck entwickelten neuen Instrumente anwenden werden.

Komprimieren, dann indexieren

Die Lösung umfasst zwei Schritte: 1) Zuerst müssen die Daten komprimiert und dann 2) indexiert werden. Die damit verbundenen Schwierigkeiten haben es in sich, geht es doch darum, die Daten zu komprimieren, in dem man sie «semantisch» strukturiert. Komprimierungsformate wie MPEG, ZIP, JPEG und, neueren Datums, JPEG2000 (siehe Kasten) sind bereits bekannt, sie sind aber zurzeit angesichts der gigantischen Datenmengen, die es zu verarbeiten gilt, noch nicht mehr als «Lösungsembryonen».

Werfen wir beispielsweise einen Blick auf die Archive des Jazzfestivals Montreux. Dabei handelt es sich mehrheitlich um Ton und Bilddaten. Die EPFL ist zurzeit mit der Archivierung dieser Daten beschäftigt. Wie soll man innert nützlicher Zeit Zugriff auf exakt jene Daten erhalten, die man sucht? Die Antwort ist in aller Munde: mittels «semantischer» Abfrage.

Die Herausforderung semantische Abfrage

Auch diese Lösung weist zahlreiche Fallstricke auf. Die Inhalte sind in diesem Zusammenhang sehr wichtig. Nun ist aber bekannt, dass die Inhalte Töne, Text, Bild und Video umfassen. Man muss also in verschiedenen Datentypen suchen. Die Suche, die gegenwärtig vorgeschlagen wird, ist unabhängig von der Art und Weise der Daten. Die Lösung heisst also Integration von Daten – erst mit integrierten Daten wird eine zielgerichtete Suche möglich.

Ein anderes Beispiel: virtuelle Sitzungen. Immer häufiger werden Sitzungen virtuell durchgeführt. Diese Tendenz wird sich angesichts der explodierenden Kosten für nicht erneuerbare Energien und damit für örtliche Verschiebungen künftig noch akzentuieren. Die Archivierung dieser Sitzungen (Politik, Wissenschaft, Verbände, Sport, Kultur) wird damit unumgänglich und verlangt nach Lösungen im Bereich Datenkompression und Lagerung/Speicherung. Entsprechende Lösungen sind zurzeit noch nicht greif bar. Über die Lösung für dieses berüchtigte «Binom der Zukunft» beugen sich heute in der ganzen Welt Heerscharen von Forschern ...

Schlussfolgerung

Die Aufgabe der Forscher hat titanische Ausmasse. Es wird noch eine gewisse Zeit dauern, bis die Berufsleute aus dem Bereich Informationsdokumentation über Instrumente verfügen werden, die es ihnen ermöglichen, ihrer Kundschaft Dienstleistungen im Bereich audiovisuelle Bestände anbieten können, die diesen Namen auch verdienen.

Anmerkung:

(1) The Expanding Digital Universe. A Forecast of Worldwide Information Growth Through 2010, unter der Leitung von John F. Gantz, März 2007.

Die Norm JPEG2000

JPEG2000 ist ein neues Bildcodierungssystem, das die modernsten Komprimierungstechniken anwendet und auf der Transformierung in Wavelets aufbaut. Die Systemarchitektur ist für eine Vielzahl von Anwendungen (von digitalen Fotoapparaten bis hin zu medizinischen Bildgebungsverfahren und anderen Schlüsselbereichen) geeignet. Die Codierung umfasst Informationen über den Inhalt sowie eine primäre Indexierung.

Pierre Vandergheynst

Professeur a? l’EPFL.