Optimizing Access to Information on the Internet

Karl Aberer
GMD-IPSI, Darmstadt
aberer@darmstadt.gmd.de
Overview

XML and eCommerce
  • Internet Business
  • Technology Platform XML
  • Information Intermediation

An economic view on global access optimization
  • Introducing the economic paradigm
  • An architecture for information commerce

Optimization of document queries for local access
  • Document query rewriting
  • Rule-based optimization for object-relational queries

Open Issues
XML and eCommerce
Business-oriented Internet Applications

Interoperability of business processes
  • Application integration, workflows, XML/EDI

Selling products and services
  • Consumer goods, procurement, financial services

Selling and delivering information goods
  • B2C: electronic books and newspapers, games, music, videos
  • B2B: news, financial information, studies and reports, software

New types of businesses: information services supporting businesses
  • auctions, recommender services, push channels, directories, search services, service chaining, portals
<table>
<thead>
<tr>
<th>Service</th>
<th>Example</th>
<th>Added value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structured search</td>
<td><a href="http://www.jobrobot.de">www.jobrobot.de</a></td>
<td>Single point of access for autonomous sources</td>
</tr>
<tr>
<td>Chaining of services</td>
<td><a href="http://www.priceline.com">www.priceline.com</a></td>
<td>Travel reservations (flight, hotel, car rental, …)</td>
</tr>
<tr>
<td>Automatic price comparison</td>
<td><a href="http://www.jango.com">www.jango.com</a></td>
<td>Comparison of prices of identical products offered by different merchants</td>
</tr>
<tr>
<td>Recommendation</td>
<td><a href="http://www.amazon.com">www.amazon.com</a></td>
<td>Tips for books based on individual profiles</td>
</tr>
<tr>
<td>Integrated portals</td>
<td><a href="http://www.tiscover.at">www.tiscover.at</a></td>
<td>Vacations in Austria (weather, hotels, events, …)</td>
</tr>
</tbody>
</table>
Problem: Access to Different Sources
Solution: Construct an Information Intermediary
Underlying Technological Platform: XML

Syntax to represent information
- generic markup language
- description of the structure of documents, data, and messages (not layout)
- definition of document types (specific markup languages)
- standardized by W3C (World Wide Web Consortium)

Architecture for (hypermedia) information processing
- XML is a family of (evolving) standards
- based on the XML syntax
- basic parts standardized by W3C

Documenttypes and XML applications
- based on XML syntax and XML architecture
- define semantics (vocabularies) and business rules (protocols)
- open or commercial
- horizontal or vertical
<?xml version="1.0" encoding="ISO-8859-1" ?>
<golfplatz id="platz0001" href="http://www.golfuehrer.de/scripts/ab2.idc?CLUBNR=0001">
  <adresse>
    <club>Golf-Club Berlin Wannsee e.V.</club>
    <strasse>Golfweg 22</strasse>
    <plz>14109</plz>
    <stadt>Berlin</stadt>
    <telefon>030-8067060</telefon>
    <fax>030-80670610</fax>
  </adresse>
  <greenfee>
    <wochentag>100</wochentag>
    <wochenende>120</wochenende>
  </greenfee>
  <platz>
    <kurzbeschreibung>Die Anlage liegt in einer idyllischen ...</kurzbeschreibung>
    <loecher>18</loecher> Golf-Club Berlin Wannsee e. v.
    <laenge>6088</laenge>
  </platz>
  <policy>
    Gäste sind willkommen. Clubausweis mit ...
  </policy>
  <handicap>
    <wochentag>34</wochentag>
    <wochenende>34</wochenende>
  </handicap>
</golfplatz>
Golfplatz Adressen Policy Plätze Handicaps

Golf-Club Berlin Wannsee e.V.

Golfweg 22 14109 Berlin 030-8067060 030-80670610
XML Architecture

Standard XML applications
- XHTML, SMIL, P3P, MathML

Specific XML Applications

Layout
- XSL
- CSS

Hyperlinks
- XLink
- XPointer

Metadata
- RDF, RDFS
- PICS

API
- DOM
- SAX

Schemas
- XSDL
- XDTL

Queries
- XSLT, XPath
- XQL, XML-QL

XML 1.0
- DTD
- Namespaces
- Unicode
- URI
XML-based Applications

Document and data exchange
- Universal file format for standard application programs
- Interoperability of business applications and DBMS
- Eventually: "the" glue for application integration

Document management
- Within an organization and for the WWW
- Cross media publishing
- Eventually: knowledge and workflow management

Electronic Commerce
- Flexible electronic data interchange (EDI)
- Interoperability and standardization of business processes
- Eventually: closing the electronic loop

Information Intermediation
- Search, reuse, and further processing of information
- Integration of information from different sources (value adding)
- Eventually: information commerce
Information Intermediary
Information Intermediation

Purpose
- bridge the gap between information providers and consumers at the *representational, extensional and operational* level

Tasks
- selecting information to increase specialization
- combining information to extend the coverage
- provide controlled access and optimize the quality of access (time)

Benefit
- reduce search cost for consumer (or producer)
  - users searching information about products
  - product information attracting user attention
- thus business opportunity: Information Commerce
Two Problems

Creating the intermediary’s information offers (global optimization)
• construct „interesting“ views from available resources
• minimize potential resource consumption and maximize potential benefit

Processing the requests against the intermediary information offers (local optimization)
• decompose the requests against the resources
• minimize actual resource consumption and maximize actual benefit
An Economic View on Global Access Optimization
Economic View on Information Intermediation

**Client layer**
- Browser

**Middle tier**
- XML Store
- Business Logic
  - (De-)composition

**Source layer**
- DBMS
- WWW-Site

---

Utility: finding information, resolving semantic heterogeneity, controlling access

Cost: physical resources consumed (time, space)

Cost: property rights required
Economic View on Information Intermediation

Utility: exposition of information to users

Cost: user attention

Cost: physical resources consumed (time, space)
Using Economic Paradigms to Optimize Resource Usage in Distributed Data Management

Economy for Managing Replicated Data
[Ferguson, Nikolaou, Yemini, 1993]

Mariposa
[Stonebraker et al, VLDB Journal 95]
Towards a Generic Architecture for Information Commerce

Functions of a Market [Bakos, Comm. of the ACM, Aug 98]

- Matching buyers and sellers
  - determination of product offers (features, aggregation)
  - search (buyers - sellers)
  - price discovery
- Facilitation of transactions
  - logistics (delivery)
  - settlement (payment)
  - trust
- Institutional Infrastructure (rules and their enforcement)
### Features of Information Product Offers

<table>
<thead>
<tr>
<th>Information type</th>
<th>Ownership rights</th>
</tr>
</thead>
<tbody>
<tr>
<td>• view/schema</td>
<td>• reselling, licensing</td>
</tr>
<tr>
<td>• metadata</td>
<td>Authentication</td>
</tr>
<tr>
<td>• types of queries supported</td>
<td>• certificates</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Access specification</th>
<th>Price model</th>
</tr>
</thead>
<tbody>
<tr>
<td>• push/pull access</td>
<td>• offer and access related components</td>
</tr>
<tr>
<td>• read/write access</td>
<td>• money vs. attention</td>
</tr>
<tr>
<td>• temporal, quantitative constraints</td>
<td></td>
</tr>
<tr>
<td>• state change based notification</td>
<td></td>
</tr>
<tr>
<td>• quality of service</td>
<td></td>
</tr>
</tbody>
</table>
Infrastructure for Information Commerce

- **consumer**
- **payment authentication**
- **producer**

**Information intermediation**
- **contract**
- **offer**

- **contract enactment protocol (requests, updates, payments)**
- **contract negotiation protocol (purchasing, negotiation, bidding)**

- **communication infrastructure message format communication protocol**
- **advertisement**
- **prospective consumer producer**
XML Application:
ICE - Information and Content Exchange Protocol

Standardized mechanism for content syndication
- automatizes establishment of syndication relationships, data transfer and result analysis
- uses XML to encode protocol messages and http as transport mechanism
- does not define content,vocabularies, mechanisms for payment, security and authentication

step 1: subscriber ➔ syndicator: ice-get-catalog
step 2: subscriber ➙ syndicator: ice-catalog
step 3: subscriber ➔ syndicator: ice-offer (select an offer)
step 4: subscriber ➙ syndicator: ice-offer (counteroffer, e.g. pull info)
step 5: subscriber ➔ syndicator: ice-offer (accept counteroffer)
step 6: subscriber ➙ syndicator: ice-subscription
step 7: subscriber ➔ syndicator: get-package, initial state ICE-INITIAL
step 8: subscriber ➙ syndicator: ice-package P1, new state

request catalog
negotiate offer/
establish contract
request/deliver
Example Revisited

For XML Broker

XSL
XQL
XML & ICE

<ice-delivery-policy
   startdate="1998-07-02T12:00:00">
   <ice-delivery-rule
      mode="pull"
      starttime="02:00:00"
      duration="P7200S"
      maxcount="1">
   </ice-delivery-rule>

   <ice-delivery-rule
      mode="push"
      url="http://xmlbroker.de/ice-in/">
      ...</ice-delivery-rule>
</ice-delivery-policy>

<www.reiseplanung.de>
   <route>
      <von>53757</von>
      <nach>93333</nach>
      <entfernung>481.9</entfernung>
      <fahrzeit>274</fahrzeit>
   </route>
</www.reiseplanung.de>

<www.wetter.de>
   <wetter>
      <plz>87724</plz>
      <datum>981001</datum>
      <temperatur>16</temperatur>
      <regen>90</regen>
      <wind>9</wind>
      <prognose>13</prognose>
   </wetter>
</www.wetter.de>

<ice-delivery-policy
   startdate="1998-07-02T12:00:00">
   <ice-delivery-rule
      mode="pull"
      starttime="02:00:00"
      duration="P7200S"
      maxcount="1">
   </ice-delivery-rule>

   <ice-delivery-rule
      mode="push"
      url="http://xmlbroker.de/ice-in/">
      ...</ice-delivery-rule>
</ice-delivery-policy>
Internet Based Infrastructure for Information Commerce

Source layer
- DBMS
- WWW-Site
- Business-application

Transformations and access layer
- Transformation
- BAPI

Middle tier
- XML Store
- Adapter
- Wrapper
- Gateway

Application
- Business Logic
- (De-)composition

Client layer
- Browser
- XML & XSL
- HTML

Payment, authentication trust centers
- OFX
- DIGSIG, P3P

WWW-Site
- Source layer
- Middle tier
- Client layer

DBMS
- SQL
- JDBC
- http
- html

XML & ICE+

Adapter
- Wrapper
- Gateway

XML

Application
- Business Logic
- (De-)composition

OFX
- DIGSIG, P3P

XML & ICE+

XML & XSL

HTML
Internal Operation of an Information Intermediary - „Development & Marketing & Sales“

Creation of new aggregate information products

- exploring the space of possible products: searching offers, create possible aggregations
- optimizing the products based on different types of context
  - analysis of available information sources (retrieval, extraction, linking)
  - analysis of consumer behavior (profile, history)
  - business knowledge
  - processing capabilities
  - statistical approaches, data mining

Solving the global optimization problem of Information Access!
Internal Operation of an Information Intermediary - „Delivery & Service“

Processing of aggregate information products
- composing data, decomposing views
- composing/decomposing the other features of the information product: coordinating access, observing ownership rights, processing payment (OPELIX)
- optimizing resource usage vs. benefit
  - e.g. cost-based query optimization

Solving the local optimization problem of Information Access!
Optimization of Document Queries
Document Query Processing

Goals
- Reduce evaluation cost
  - simplifying the query
  - e.g. minimizing number of navigations in DOM-like representation through changing navigation pattern
  - exploiting indices on the document structure
- Enable the mapping to imported views

Approach
- Algebraic representation of the query
  - successful approach for relational queries
  - determine document query algebra and equivalence rules
- Semantic optimization
  - schemas not mandatory for XML documents
  - structure of DTDs can be additionally exploited for modifying the query
Document Query Algebra and Rules

Document query algebra

- A, B document element sets
- \( A \cap B = \{ a \in A \mid b \in B \text{ such that } b \text{ contained in } a \text{ in the document tree} \} \)
- \( B \cup A = \{ b \in B \mid a \in A \text{ such that } a \text{ contained in } b \text{ in the document tree} \} \)
- set operators \( \cap, \cup \)
- selection operators \( \text{pred} \)

Rules

- \( A \cap (B \cup C) \Rightarrow (A \cap B) \cup C \)
- \( A \cap (B \cap C) \Rightarrow (A \cap B) \cap C \)
- \( A \cap (B \cap C) \Rightarrow (A \cap B) \cap C \) ?
- \( (A \cap B) \cup C \Rightarrow A \cap (B \cap C) \) ?

Diagram:
XQL query:
//Vacation[//Price<1000]//Flight/Time

Algebraic transformations:

Time (Flight (Vacation \text{\(p<1000\)} (Price)))

\iff (\text{entrance location: all paths pass Arrangement})

Time (Flight (Arrangement (Vacation \text{\(p<1000\)} (Price))))

\iff (\text{exclusivity: Flight always contained in Arrangement})

Time (Arrangement (Vacation \text{\(p<1000\)} (Price))))

\iff (\text{Index for Time-Arrangement inclusion relationship})

INDEX \text{\(\text{Time,Arrangement}\)} (Arrangement (Vacation \text{\(p<1000\)} (Price)))
Results on XML Query Rewriting

- Algebra and general document query rewrite rules described
- DTD properties for semantic rewriting (exclusivity, obligation, entrance location) and algorithms for their determination [Böhm, Aberer, Öszu, Gayer, ADL 98]
- Deterministic heuristic rewrite strategy to systematically simplify queries and enable index match by semantically rewriting the query [Che, Aberer, IDEAS 99]
Cost-based Document Query Optimization

Assumption
- documents are stored in an *object-relational* DBMS, e.g. in a DOM like fashion

Goal
- extend existing relational query processing approaches

Approach
- rule-based, cost-based query optimizer
- existing optimizer implementation framework (Volcano)

Representing XML in DOM like fashion
[Böhm, Aberer, Neuhold, Yang, VLDB J. 97]
[Böhm, Aberer, Klas, MTAP 99]
Rule-based Optimization of OR Queries

Algebra rules
- support an extended relational algebra
  \[
  \text{map}_{a=f(ai)}(S) = \{ [a_1:v_1,\ldots,a_n:v_n,a: f(v)] \mid [a_1:v_1,\ldots,a_n:v_n] \in S \}\]
- provide a rule-based implementation to allow flexible addition of rules to the standard relational rules [Aberer, Fischer, DE95]
- allows to express additional rules for document semantics
- problem: uncontrolled generation of alternative expressions

Known results on join enumeration
- join enumeration used in approach for cost-based optimization using dynamic programming
  - complexity $O(k^3)$ (Starburst), $k$ number of join predicates
- efficient rule based join enumeration
  - duplicate-free generation of join orders for non-cyclic join queries possible [Pellenkoft et al, VLDB97]
Arrangement[ Price[ !text()="1000" ] && Dest[ !text()="Lausanne" ] ] (XQL notation)
Approach and Result

Approach

- design the rule system, such that each rule application "simplifies" the expression
- thus duplicates are avoided!
- redesign/experiment till all desired expressions are generated (simulator implemented in Mathematica)

Result

- proven that it is possible for linear select-join queries with complexity $O(k^3)$, $k$ number of joins
- duplicates almost completely avoided
  [Aberer, Che, Böhm DOOD97]
Cost-based Query Optimization within the Economic Framework

New requirements

- cost aggregation may exhibit „nonstandard“ behavior
- query processing time is relevant
- optimal solution not necessarily needed
- access policies may lead to new types of optimization problems (e.g. combined push/pull)
Issues in Information Commerce

Consolidated architecture
Generic tools for processing aggregate information products
• formal models describing composition of data, access policies, ownership rights and payment
• run-time optimization and design-time optimization driven by economic principles

Scope of markets/granularity of information products
• global, national, regional, organizational, system

Interdisciplinary work required
• economy, agents, networks, information, security

Automatization of information business
• compare to developments on financial markets
The end

Thank you for your attention!