Chapter 1

Introduction

The World Wide Web is a global information space consisting of information from a multitude of autonomous information providers [JW04]. Web-based information systems provide access to this information space. They integrate information from multiple information providers and present integrated information to their users.

The key success factor of the Web is the vast amount of web-accessible information. On the other hand, its openness and the autonomy of information providers make the Web vulnerable to inaccurate, misleading, or outdated information. Information quality problems arise in various application domains of web-based information systems:

Search Engines provide access to billions of web documents and an increasing number of structured information sources [Mil05]. The quality of provided information varies widely and the huge amount of accessible information obscures relevant information.

News Portals like Google News[1] or NewsNow[2] aggregate news articles from a wide range of newspapers and news agencies and assemble them according to user's interests. Besides classic newspapers and news agencies, an increasing number of private or company-run weblogs have sprung up and offer news on a multitude of topics [Wik06a]. These weblogs can be accessed through news feed aggregators like Technorati[3] which offers access to 49.7 million weblogs (August 2006) [Tec06]. As news providers have different views of the world and different levels of knowledge, news may be inaccurate or biased.

[1] http://news.google.com/ (retrieved 09/25/2006)

[2] http://www.newsnow.co.uk/ (retrieved 09/25/2006)

[3] http://technorati.com/ (retrieved 09/25/2006)

Financial Information Portals like Wall Street Journal Online[4] or Yahoo Finance[5] integrate stock quotes, financial news, company profiles, analyst reports from multiple information sources and provide discussion forums about investment related topics. The expertise of information providers about specific markets and companies varies widely. Therefore, judgments from certain sources are more accurate than others and investors are confronted with conflicting advice.

Electronic Markets and Bargain Finder Services integrate information about products, product offers, and merchants from multiple sources. Especially business-to-consumer and consumer-to-consumer markets like eBay[6] suffer from scams by sellers who publish false offers, try to build up a misleading reputation through faked transactions or boost their sales through biased product reviews.

Knowledge Management Systems and employee portals enable information consumers to access a multitude of information sources from inside an organization, external information providers, partner organizations, and the Web. A recent trend in knowledge management is the success of Wiki systems, which enable communities of knowledge providers to collaboratively author information [Wik06b]. As some information providers may be experts on a certain topic while others are less informed, it is crucial for knowledge management systems to provide means to distinguish high quality from low quality information.

Online Communities like MySpace[7], Del.icio.us[8], Flickr[9], or YouTube[10] are used by large numbers of information providers to share information. MySpace for instance has accumulated 67 million members (March 2006) [KS06] who use the service to share information about personal interests as well as music, photos, and videos. The quality of provided information varies widely, and again the amount of accessible information blurs relevant information.

Web 2.0 Mashups. Major web data sources like Google, Amazon, eBay, Technorati, Flickr, and YouTube have started to provide public query interfaces to their databases. The availability of these Web APIs has

[4] http://online.wsj.com/ (retrieved 09/25/2006)

[5] http://finance.yahoo.com/ (retrieved 09/25/2006)

[6] http://www.ebay.com (retrieved 09/25/2006)

[7] http://www.myspace.com/ (retrieved 09/25/2006)

[8] http://del.icio.us/ (retrieved 09/25/2006)

[9] http://www.flickr.com/ (retrieved 09/25/2006)

[10] http://www.youtube.com/ (retrieved 09/25/2006)

set off the development of Web 2.0 mashups, i.e. web-based information systems that integrate information from multiple Web APIs. An example of a mashup is Virtual Places[11], an application that renders different types of location-related information such as photos, books and websites about a place, together on a map. Again, the quality of information provided by the Web APIs may vary and the huge amount of accessible information blurs relevant information.

Semantic Web Applications. The Semantic Web [Her06] is a global information space consisting of linked data [BL06a]. Semantic Web data sources publish information on the Web using the Resource Description Framework (RDF) [HSB06] as shared data model. Examples of Semantic Web data sources are LiveJournal[12], Geonames[13], Dbpedia[14], and the RDF Book Mashup[15]. Semantic Web applications enable users to navigate and query linked data on the Web. Examples of Semantic Web applications are the Tabulator browser [BL06b] and the doap:store search engine[16]. Assuring information quality is problematic within Semantic Web applications as they operate on an unbound, dynamic set of autonomous data sources.

Shielding information consumers from low quality information is increasingly recognized as a crucial aspect in the design of web-based information systems because of the problems outlined above.

1.1 Problem Definition

Web-based information systems provide access to information originating from multiple information providers. Information providers have different levels of knowledge, different views of the world, and different intentions. Therefore, provided information may be wrong, biased, and inconsistent.

Before information from the Web is used to accomplish a specific task, its quality should be assessed according to task-specific criteria. Based on the assessment result, information may be accepted or rejected for a specific task.

[11] http://apps.nikhilk.net/VirtualPlaces/ (retrieved 09/25/2006)

[12] http://www.livejournal.com/ (retrieved 09/25/2006)

[13] http://www.geonames.org/ (retrieved 09/25/2006)

[14] http://dbpedia.org/docs/ (retrieved 09/25/2006)

[15] http://sites.wiwiss.fu-berlin.de/suhl/bizer/bookmashup/ (retrieved 09/25/2006)

[16] http://doapstore.org/ (retrieved 09/25/2006)

In everyday life, we use a wide range of different policies to assess the quality of information: We might accept information from a friend on restaurants, but distrust him on computers; regard scientific papers only as relevant, if they have been published within the last two years; or believe foreign news only when they are reported by several independent sources. Which policy is chosen depends on the specific task at hand, our subjective preferences, and the availability of information quality-related meta-information, such as ratings or background information about information providers.

This thesis introduces an innovative solution to quality-driven information filtering in the context of web-based information systems. Instead of having the designer of an information system decide for the user on a single, fixed method to assess the quality of information, the user is empowered to employ a similar wide range of filtering policies as she is using in the off-line world.

This user-centric approach to information filtering is explored by developing a policy-based information filtering framework. The framework can be used within web-based information systems to enable information consumers to employ task-specific filtering policies. The thesis divides the general problem of policy-based information filtering into three subproblems:

Representing Meta-Information. As a prerequisite for being able to employ different quality-based filtering policies, it is necessary to represent information together with quality-related meta-information using an adequate data model.

Expressing Information Filtering Policies. In order to enable information systems to filter information according to the user's task-specific quality requirements, a flexible language for expressing filtering policies is needed.

Explaining Filtering Decisions. The key factor for a user to trust filtering decisions is her understanding of the filtering process. In order to facilitate the user's understanding of this process, information systems should provide explanations why information satisfies a given policy.

Each of these problems is addressed and solved during the course of this thesis.

1.2 Thesis Outline

The thesis is divided into three parts. The first part introduces the concept of information quality and gives an overview about different metrics that can be used to assess information quality in the context of web-based information systems. The second part proposes a data model for representing information from the Web together with quality-related meta-information. The third part of the thesis develops a language for expressing quality-based information filtering policies and describes its implementation as part of the WIQA framework.

In the following, the thesis is outlined by summarizing each chapter.

Chapter 1: Introduction. This chapter motivates and states the problem of quality-based information filtering in the context of web-based information systems. Web-based systems are susceptible to minor quality information as they integrate information from multiple autonomous information providers. Therefore, it is necessary to assess the quality of information before it is used to accomplish a specific task. The general problem of policy-based information filtering has been divided into three subproblems which will be solved in the course of this thesis.

PART I: Information Quality and the Web

Chapter 2: Information Quality. Information quality is commonly seen as the fitness for use of information. It is a multidimensional construct as the fitness for use may depend on various factors. This chapter discusses the different information quality dimensions.

Chapter 3: Information Quality Assessment. Information quality assessment is the process of evaluating if a piece of information meets the information consumer's quality requirements for a specific task. Information consumers can use a wide range of different metrics to assess information quality dimensions. This chapter gives an overview about the different metrics and examines their applicability in the context of web-based systems. The metrics are classified into three categories: Content-, context-, and rating-based metrics. Afterwards, the concept of quality-based information filtering policies is introduced.

PART II: Representation of Meta-Information. As a prerequisite for being able to employ different quality-based information filtering policies, it is necessary to represent information together with quality-related meta-information using an adequate data model. This part of the thesis proposes a data model for representing information from the Web together with quality-related meta-information.

Chapter 4: The Resource Description Framework. This chapter examines whether the Resource Description Framework (RDF) data model [KC04], a state-of-the-art data model for web-based information systems, satisfies the requirements arising from information quality assessment. It is concluded that the RDF reification mechanism is not adequate for representing information together with quality-related meta-information.

Chapter 5: Named Graphs. This chapter extends the RDF data model to the Named Graphs data model in order to eliminate the shortcomings found in the previous chapter. Afterwards, the TriG and TriX syntaxes for exchanging sets of named graphs are introduced.

Chapter 6: The Semantic Web Publishing Vocabulary. An important type of quality-related meta-information is provenance information. This chapter develops the Semantic Web Publishing vocabulary for representing provenance information and for assuring the origin of information with digital signatures.

Chapter 7: Use Case: Financial Information Integration. This chapter applies the Named Graphs data model and the Semantic Web Publishing Vocabulary within a financial information integration scenario. The data model and the vocabulary are used to represent information about companies, stocks, financial news, analyst reports, and postings from investment related discussion forums, together with quality-related meta-information. This scenario will be used as a running example throughout the proceeding chapters.

Chapter 8: Summary. This chapter summarizes the proposed solution for representing information from the Web together with quality-related meta-information.

PART III: The WIQA Framework. The third part of the thesis describes WIQA Information Quality Assessment Framework. The WIQA framework is a set of software components that can be employed by web-based information systems to enable users to filter information using a wide range of different filtering policies.

Chapter 9: Expressing Information Filtering Policies. This chapter develops WIQA-PL, a policy language for expressing quality-based information filtering policies. WIQA-PL policies may combine content-, context- and rating-based assessment metrics.

Chapter 10: Explaining Assessment Results. The key factor for a user to trust filtering decisions is her understanding why information fulfills a given policy. This chapter describes the capabilities of the WIQA framework to explain filtering decisions.

Chapter 11: Implementation. This chapter describes the implementation of the WIQA framework. The implementation consists of two parts: The NG4J - Named Graph API for Jena, a general purpose extension to the Jena Semantic Web framework [CDD$^+$04] for handling sets of named graphs, and the WIQA Filtering and Explanation Engine which enables applications to filter a set of named graphs using WIQA-PL filtering policies.

Chapter 12: The WIQA Browser is an example application which uses the WIQA framework. The browser demonstrates how information quality filtering capabilities can be integrated into a standard web browser. The browser enables users to collect structured information from web pages. Collected information is stored in a local repository together with quality-related meta-information using the Named Graphs data model. Users can filter the content of the repository using WIQA-PL filtering policies and can retrieve explanations about filtering decisions.

Chapter 13: Related Work. This chapter compares the WIQA framework with related work.

Chapter 14: Conclusion. The final chapter summarizes the contributions of this thesis and outlines directions for future research.

1.3 Research Method

“Two paradigms characterize much of the research in the Information Systems discipline: behavioral science and design science. The behavioral-science paradigm seeks to develop and verify theories that explain or predict human or organizational behavior. The design-science paradigm seeks to extend the boundaries of human and organizational capabilities by creating new and innovative artifacts” [HMPR04]. “IT artifacts are broadly defined as constructs (vocabulary and symbols), models (abstractions and representations), methods (algorithms and practices), and instantiations (implemented and prototype systems)” [HMPR04]. Artifacts are evaluated with respect to the utility provided in solving the problems they were designed for. An artifact must be innovative, solving a heretofore unsolved problem or solving a known problem in a more effective or efficient manner.

This thesis uses the design-science research method. Chapters 2 and 3 analyze the problem of quality-driven information filtering in the context of web-based systems by reviewing related research work and experience from deployed information systems. This analysis is taken as a basis for deriving requirements for a general, policy-based information filtering framework. Afterwards, three artifacts are designed which together fulfill the requirements: The Named Graphs data model, the Semantic Web Publishing Vocabulary and the WIQA-PL policy language. The novelty of the artifacts is shown by comparing them with related approaches. Evidence for their utility to solve the described problem is provided by implementing them as a proof-of-concept in the form of the WIQA framework and by applying the framework within a financial information integration scenario.

1.4 Published Work

Parts of the work presented in this thesis have been published in international journals and the proceedings of international conferences and refereed workshops. Publications relating to this work are listed below:

International Journals

- Jeremy Carroll, Christian Bizer, Patrick Hayes, Patrick Stickler: Named Graphs. Journal of Web Semantics, volume 3(4), pages 247-267, 2005.
- Robert Tolksdorf, Christian Bizer, Rainer Eckstein, Ralf Heese: Trustable B2C Markets on the Semantic Web. International Journal of Computer Systems Science & Engineering, volume 19(3), pages 199-206, 2004.

International Conferences

- Jeremy Carroll, Christian Bizer, Patrick Hayes, Patrick Stickler: Named Graphs, Provenance and Trust. In Proceedings of the 14th International World Wide Web Conference, pages 613 - 622, Chiba, Japan, May 2005.
- Christian Bizer, Richard Cyganiak, Rowland Watkins: NG4J - Named Graphs API for Jena. In 2nd European Semantic Web Conference - Posters and Demonstrations, Heraklion, Greece, 2005.

- Christian Bizer, Radoslaw Oldakowski: Using Context- and Content-Based Trust Policies on the Semantic Web. In 13th World Wide Web Conference - Posters and Demonstrations, New York, USA, 2004.

Refereed Workshops

- Christian Bizer, Richard Cyganiak, Tobias Gauss, and Oliver Maresch: The TriQL.P Browser: Filtering Information using Context-, Content- and Rating-Based Trust Policies. In Proceedings of the Semantic Web and Policy Workshop at the 4th International Semantic Web Conference, Galway, Ireland, 2005.

- Jeremy Carroll, Christian Bizer, Patrick Hayes, Patrick Stickler: Semantic Web Publishing using Named Graphs. In Proceedings of the Workshop on Trust, Security, and Reputation on the Semantic Web at the 3rd International Semantic Web Conference, Hiroshima, Japan, 2004.