Fortschritt-Berichte VDI

VDI

Reihe 10

Informatik/
Kommunikation

Dipl.-Inf. Jan Werrmann,
Stuttgart

Nr. 849

AIRS - Advanced Ontology-based Information Retrieval System

AIRS – Advanced Ontology-based Information Retrieval System

Dissertation
zur Erlangung des akademischen Grades
DOKTOR RER. NAT.

der Fakultät für
Mathematik und Informatik
der FernUniversität
in Hagen

von
Jan Werrmann
geb. in Borna

Hagen 2016

Fortschritt-Berichte VDI

Reihe 10

Informatik/
Kommunikation

Dipl.-Inf. Jan Werrmann,
Stuttgart

Nr. 849

AIRS - Advanced Ontology-based Information Retrieval System

VDI verlag

Werrmann, Jan
AIRS – Advanced Ontology-based Information Retrieval System
Fortschr.-Ber. VDI Reihe 10 Nr. 849. Düsseldorf: VDI Verlag 2016.
174 Seiten, 49 Bilder, 5 Tabellen.
ISBN 978-3-18-384910-9, ISSN 0178-9627,
€ 62,00/VDI-Mitgliederpreis € 55,80.

Keywords: Information Retrieval – Ontology Development – Knowledge Representation – Advanced Search Technologies – Heterogeneous Document Landscapes

Obtaining the right information at the right time is one of the main challenges for modern societies. This holds especially for companies that must handle complex business processes that require case dependent information. Unfortunately, case dependent and relevant information is often widespread over different document systems. Users must interact with various applications and search for semantically related (and helpful) documents without any, or with only little, support by the disparate retrieval systems. In this work, a system called Advanced ontology-based Information Retrieval System (AIRS) is introduced that includes methods of state-of-the-art enterprise search technology and combines them with an ontology called AIRS Knowledge Base (AIRSKB). AIRS is deeply integrated with advanced information retrieval technologies to make search processes in large heterogeneous document landscapes more effective and increase the quality of search results.

Bibliographische Information der Deutschen Bibliothek
Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliographie; detaillierte bibliographische Daten sind im Internet unter http://dnb.ddb.de abrufbar.

Bibliographic information published by the Deutsche Bibliothek
(German National Library)
The Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliographie (German National Bibliography); detailed bibliographic data is available via Internet at http://dnb.ddb.de.

ISSN 0178-9627
ISBN 978-3-18-384910-9

Acknowledgments

The following work was developed during my employment at the Global Service & Parts (GSP) department of Daimler AG. Daimler started a research project for the optimization of workshop literature access and I am thankful that Daimler gave me the opportunity to participate in it.

I came in touch with Professor Bernd J. Krämer from FernUniversität in Hagen who later became my supervisor. I would like to thank Professor Krämer very much for his support from the start of my research project until the end in all facets. He encouraged me to add a collective intelligence approach into my research project.

I also would like to thank Professor Gerhard Heyer from Universität Leipzig where I received my degree in computer science. He was the supervisor of my diploma thesis and through him I came in touch with Daimler. At Daimler, I would like to thank all my colleagues who supported this work.

Very special thanks to my family who supported me in writing my thesis. Especially to my wife Natalie Werrmann and to my parents Dr. Angela Werrmann and Udo Werrmann who gave me the motivation and strength. Last but not least, I want to thank all my friends for just giving me time to finish my thesis.

For my grandfather Johannes Ludwig.

Thank you for your math lessons when I was a lazy child ...

Contents

Abstract

Obtaining the right information at the right time is one of the main challenges for modern societies. This holds especially for companies that must handle complex business processes. These business processes require case dependent information. But relevant information is often widespread over different document systems. Enterprise search is a field of research that focuses on the challenges of information access. Unfortunately, a deep integration of knowledge networks as well as relationships between index documents is not sufficiently supported by enterprise search systems. Often, good retrieval results depend on these relationships, because the documents of the systems correlate somehow to each other regarding a special business case.

This presents a heterogeneous information system and document landscape to the employees who must find the right piece of information they need from different retrieval systems. In the end, an employee must interact with various desktop applications and search for semantically related (and helpful) documents without any, or with only little, support by the disparate retrieval systems.

The main motivation for this work can be summarized in the following research questions:

1. Can a single systems view be provided for all of the case-related documents kept in different retrieval systems?

2. Can seamless and guided access across these disparate and disconnected retrieval systems be designed?

3. Can the quality of retrieval results and the effectiveness of the retrieval process be improved by exploiting user feedback?

4. Can a technical solution be developed that is accepted by users in the field?

To address these research questions, enterprise search technologies were combined with knowledge representation techniques based on ontologies and user feedback processing. For this approach, a system called Advanced ontology-based Information Retrieval System (AIRS) was developed. It includes methods of state-of-the-art enterprise search technology and combines them with an ontology called AIRS Knowledge Base (AIRSKB). AIRSKB provides an overarching knowledge structure modeling documents, document sources and explicit or deduced relationships between them. This knowledge structure now represents a homogeneous and coherent search space. It is deeply integrated with advanced information retrieval technologies to make search processes in large heterogeneous document landscapes more effective and increase the quality of search results. AIRSKB also serves to capture the collective intelligence of knowledge producers and knowledge users, i.e., employees. To demonstrate the feasibility of the approach and evaluate its innovative capabilities and its usability, a prototype system, called AIRS Prototype, was designed, implemented, and tested in a domain of maintenance, service and repair of cars in car workshops. These field tests included:

1. A comparison of current workshop retrieval systems with AIRS Prototype and
2. system tests along predefined domain-specific test scenarios.

The field tests were carried out with workshop employees exhibiting many years of professional experience in workshop services. They showed that the new ontology-based retrieval is superior to the existing retrieval technology and that the collective feedback of workshop experts enables the automatic and valid reconstruction of hidden document relationships.

Zusammenfassung

Die richtigen Informationen zum richtigen Zeitpunkt zu bekommen ist eine der größten Herausforderungen moderner Gesellschaften. Das gilt speziell für Firmen, die mit komplexen Prozessen hantieren müssen. Diese Prozesse verlangen Fall-abhängige Informationen für spezielle Arbeitsschritte. Allerdings sind die Informationen oft über verschiedene Dokumentensysteme verteilt. Enterprise Search ist ein Forschungsgebiet, das sich auf die Herausforderung des Informationszugangs spezialisiert hat. Leider wird die tiefe Integration von Wissensnetzwerken oder Relationen zwischen Indexdokumenten von Enterprise Search-Systemen nur ungenügend unterstützt. Oft hängen gute Suchergebnisse von diesen Relationen ab, weil Dokumente der Systeme im Bezug zu einem Anwendungsfall miteinander in Verbindung stehen.

Insgesamt gesehen stellt sich für Mitarbeiter von Firmen oft eine heterogene System- und Dokumentenlandschaft dar, in der nach erforderlichen Informationen in verschiedenen Systemen recherchiert werden muss. Das Resultat ist, dass der Mitarbeiter mit verschiedenen Applikationen hantieren und nach semantisch verbundenen (und hilfreichen) Dokumenten ohne oder mit wenig Unterstützung der unterschiedlichen Recherchesysteme suchen muss.

Die hauptsächliche Motivation für diese Arbeit kann in den folgenden Forschungsfragen zusammengefasst werden:

1. Kann eine einheitliche Sicht auf fallrelevante Dokumente der verschiedenen Recherchesysteme hergestellt werden?

2. Kann ein nahtloser Zugriff auf diese ungleichen und nicht verbundenen Recherchesysteme bereitgestellt werden?

3. Wie kann die Qualität der Rechercheergebnisse und die Wirksamkeit der Recherchesysteme durch die Verwendung des Feedbacks von Systemnutzern verbessert werden?

4. Kann eine prototypische Lösung entwickelt werden, die Akzeptanz bei den Systemnutzern findet?

Um diese Forschungsfragen zu beantworten, wurden Enterprise-Suchtechnologien mit Techniken der Wissensrepräsentation (basierend auf Ontologien) und automatischer Feedback-Verarbeitung verbunden. Für den Ansatz wurde das System Advanced ontology-based Information Retrieval System (AIRS) entwickelt, das Methoden von Enterprise-Suchtechnologien nach aktuellem Stand der Technik verwendet und mit einer Ontologie, AIRS Knowledge Base (AIRSKB), verbindet. AIRSKB stellt eine übergreifende Wissensstruktur dar, welche Dokumente, Quellen sowie feste und adaptive Relationen zwischen den Dokumenten und Quellen beinhaltet. Dadurch fungiert die Wissensstruktur als homogener und kohärenter Suchraum. Sie ist tief integriert in erweiterte Information-Retrieval-Technologien um Suchprozesse in großen heterogenen Dokumentenlandschaften effektiver zu gestalten und die Qualität der Suchergebnisse zu verbessern. AIRSKB dient auch dazu,

die kollektive Intelligenz der Wissensproduzenten und Wissensnutzer, d.h. der Mitarbeiter zu erfassen. Um die Durchführbarkeit des Ansatzes zu demonstrieren und seine Innovationskraft und Benutzerfreundlichkeit zu bewerten, wurde ein Prototyp, AIRS-Prototype, entwickelt und in der komplexen Business-Domäne von Service, Wartung und Reparatur von Fahrzeugen in Kfz-Werkstätten verprobt. Diese Feldtests beinhalteten:

1. Einen Vergleich der Dokumentenrecherche in aktuell existierenden Werkstattrecherchesystemen mit der Dokumentenrecherche von AIRS-Prototype und

2. Systemtests, basierend auf vordefinierten domänenspezifischen Test-Szenarien.

Die Feldtests wurden mit Werkstattmitarbeitern durchgeführt, die jeweils über langjährige Berufserfahrung im Werkstattumfeld verfügten. Die Tests zeigten, dass der ontologiebasierte Suchansatz bessere Ergebnisse erzielt als die existierenden Recherchesysteme. Ebenfalls zeigte sich, dass kollektives Feedback der Werkstattexperten es ermöglicht, versteckte Dokumentenbeziehungen automatisch und valide zu rekonstruieren.