- Planung und Steuerung: Recherchebericht -

MSP-13 - Integration eines Semantischen
Tagging Systems in Microsoft Sharepoint

Version: 0.3

Projektbezeichnung	MSP-13 - Integration eines Semantischen Tagging Systems in Microsoft Sharepoint	
Projektleiter	Martin John	
Verantwortlich	Recherche	
Erstellt am	07.01.2013	
Zuletzt geändert	21:50 07.01.13	
Bearbeitungszustand	X	in Bearbeitung vorgelegt fertig gestellt
Dokumentablage	GIT/Dokumentation/RecherchebberichtVersion0.1.pdf	
V-Modell-XT Version	1.4	

Inhaltsverzeichnis

1. Überblick

Mit Hilfe eines existierenden Knowledge Extraktion Frameworks (z.B. AKSW Fox) sollen Dokumente in einem Microsoft Sharepoint Server automatisch getaggt werden. Dies im Hintergrund geschehen und die erzeugten Metadaten sollen abgespeichert und angezeigt werden können. Die Integration mit dem Sharepoint Server soll auf typische Weise erfolgen

- Einrichten eines SharePoint Server
- Erstellen von Testdaten
- Automatisches Auslesen des Textes (im Hintergrund)
- Generierung NIF
- Übergabe der Daten in FOX
- Speicherung der Metadaten
- Visualisierung der Metadaten

2. Begriffe

2.1 Sharepoint

Sharepoint ist ein Produkt von Microsoft welches die Zusammenarbeit von Projektteams-koordinieren soll. Die Vorteile sind eine gemeinsame Plattform, die Zusammenarbeit und die schnelle Reaktionsfähigkeit auf die Businessanforderungen. Es ist Dokumentenmanagementsystem und Informationsmanagementsystem (Webseiten mit einem intuitiven CMS). Damit ist die Plattform anpassbar auf die Prozesse und Strukturen von Unternehmen, Gruppen und Projekten. Es kann Daten von verschiedenen Formaten und Herstellern über eine einzige Schnittstelle verfügbar machen. Es gibt die Varianten "Sharepoint Foundation" mit den Grundfunktionalitäten und "Sharepoint Server". Sharepoint Server ist die portalbasierte Version von Microsoft Sharepoint. Eine Sharepoint Server Umgebung verfügt über mindestens einen Sharepoint Foundation Server. Zusätzlich zu den Funktionen der Foundation bietet Server weitere Dienstanwendungen und spezielle Funktionen.

2.2 Knowledge Extraction Framework

Ist die Schaffung von Wissen aus strukturierten (relationalen Datenbanken, XML) und unstrukturierten (Text, Dokumente, Bilder) Quellen.

FOX erstellt RDF Tripel. Es ist ein Java Framework.

2.3 CMS

Ein Content-Management-System ist eine Software zur gemeinschaftlichen Erstellung, Bearbeitung und Organisation von Inhalten (Content) zumeist in Webseiten, aber auch in anderen Medienformen.

2.4 RDF

Resource Description Framework - ermöglicht die Beschreibung von Web-Ressourcen, wobei es auf XML basiert. RDF dient der Darstellung von Internetdaten, und ist dabei auch für Computer verständlich. RDF kann bei der Bildung von Semantic Web angewandt werden. RDF ist ein Tripel, dh.: zur Darstellung von einem Ressource braucht man einen Subjekt, Prädikat und Objekt.

2.5 SPARQL

SPARQL ist eine graph-basierte Anfragesprache für RDF.

2.6 Tagging

Ist eine Kategorisierung von Informationen.

Tagging System

Ein System, dass die Art und Weise des Taggings organisiert und die Verwaltung und Nutzung der Tags ermöglicht. Dabei gibt das System vor, ob das Tagging mittels freiem Vokabular oder mittels einer Auswahl an möglichen Tags durchgeführt wird.

2.7 Metadaten

Sind strukturierte Daten, die Informationen über andere Informationsressourcen enthalten.

2.8 Framework

Ein Framework ist ein Programmiergerüst, dass in der Softwaretechnik, insbesondere im Rahmen der objektorientierten Softwareentwicklung sowie bei komponentenbasierten Entwicklungsansätzen, benutzt wird.

2.9 NIF

Ist ein RDF/OWL[1]-basiertes Format, dass die Arbeit zwischen Natural Language Processing (NLP) Tools, Sprachressourcen und Anmerkungen erreichen will.

2.10 NLP

Natural Languag Processing ist ein Bereich der Informatik, Künstliche Intelligenz und Linguistik mit den Wechselwirkungen zwischen Computern und Menschen (natürlich) betreffenden Sprachen.

2.11 C#

C# ist eine von Microsoft im Rahmen von .NET entwickelte Sprache. Das .NET-Framework ist eine Plattform bestehend aus einer Laufzeitumgebung, in der die Programme ausgeführt werden und einer Sammlung von Klassenbibliotheken und Schnittstellen. Hauptsächlich ist C# objektorientiert.

2.12 Server

Ist ein Software Programm, dass mit anderen Programmen kommuniziert (Bsp. Clients, Server etc.) um bestimmte Dienstleistungen anzubieten.

[1] OWL: Web Onthology Language, Beschreibungssprache nach der Syntax von RDF.

2.13 RDF-Graph-Modell

Die grundlegende Struktur des RDF ist eine Ansammlung von 'Triplen'. Jedes Triple besteht aus einem Subjekt, Prädikat (auch Eigenschaft) und Objekt. Der Verbund mehrerer solcher Triple heißt RDF Graph, welcher mittels Knoten (entsprechen Subjekt oder Objekt) und Kanten (repräsentieren Prädikate) als gerichteter Graph dargestellt werden kann. Jedes Triple gibt Auskunft über ein Relation (Prädikat) zwischen zwei Dingen die durch Subjekt und Objekt beschrieben werden. In natürlicher Sprache würde sich ein Triple in etwa so lesen: Subjekt X hat die Eigenschaft (Prädiket) z, deren Wert/Aussage gleich Objekt Y ist. Die Aussage eines RDF Graphen setzt sich aus den Aussagen jedes seiner Triple zusammen (mittels log. AND-Verknüpfung).

Ein Knoten eines RDF Graphen kann eine URI-Reference, eine URI, ein Literal oder 'Blank' sein. Kanten (Eigenschaften) sind URI-Referenzen welche die Relation zwischen Subjekt und Objekt beschreiben. Eine URI-Referenz, bzw. ein Literal welches als Knoten verwendet wird gibt an, was dieser Knoten repräsentiert.

2.14 URI / URL / URN

Uniform Ressource Identifier, kurz URI, sind Zeichenfolgen, welche beliebige Ressourcen (Webseiten, Bilder, etc.) aus dem Web mittels eindeutiger Adressen identifiziert. URIs können als Zeichenfolge in digitale Dokumente, z.B. HTML oder PHP eingebunden werden.

URLs und URNs sind dabei Unterarten von URIs.

Uniform Ressource Locators (URLs) benennen eine Ressource über ihren primären Zugriffsmechanismus wie http oder ftp. Unified Ressource Names (URNs) identifizieren eine Ressource mittels eines vorhandenen oder frei zu vergebenden Namens, z. B. urn:isbn.

2.15 URI-Referenz

Eine URI-Referenz ist eine Zeichenfolge, die eine URI und somit die Ressource dieser URI repräsentiert. URI-Referenzen werden zu vollwertigen URIs konvertiert, indem sie anhand einer Basis-URI nach einem bestimmten Algorithmus aufgelöst werden. (wikipedia)

2.16 Literale

Literale geben die Werte einer Aussage/Eigenschaft (z.B. Name, Zahlen, Datum usw.) als String an. Jeder Wert eines Literals könnte auch als URI angegeben werden, doch ist es oftmals einfacher und anschaulicher ein Literal zu verwenden.

Man unterscheidet:

- 'plain literals' – normaler String mit einer Aussage
- 'typed literals' – String mit angefügter URI-Referenz zu einem Datentyp (Wertangabe)
- Ein Literal kann das Objekt eines RDF-Ausdrucks sein, nie aber ein Subjekt oder Prädikat.

2.17 Entailment

Bedeutungen und Folgerungen innerhalb eines RDF-Graphen wird durch das formale Konzept des Entailment (Folgebeziehung) gestützt.

Einfach ausgedrückt: Ausdruck A hat einen anderen Ausdruck B zur Folge, falls jede Bedingung die A 'wahr' macht auch B 'wahr' macht.

Auf dieser Grundlage kann man die 'Wahrheit' von B folgern, wenn man von der 'Wahrheit' von A ausgeht oder dies bereits gezeigt hat.

2.18 Blanks

Ein 'Blank'-Knoten besteht weder aus URI noch aus einem Literal und wird in der RDF-Syntax als UNIQUE-Knoten verwendet, welcher in mehreren RDF-Statements verwendet werden kann, aber keinen eindeutigen Namen hat.

3. Konzepte

3.1 RDF

Das RDF (Ressource Description Framework) bezeichnet eine Familie von Standards des W3C (World Wide Web Consortiums) zur formalen Beschreibung von Informationen über Objekte. Dabei war es ursprünglich zur Beschreibung von Metadaten aus dem World Wide Web entwickelt worden. Jedoch wird RDF heute auch für andere Anwendungen verwendet. So zum Beispiel für Katalogdienste und der allgemeinen Wissensrepräsentation. Dabei ist RDF eine Kernkomponente des Semantischen Webs um die Prinzipien des WWW (Verknüpfung, Offenheit, Heterogenität) von Dokumenten auf allgemeine Daten zu übertragen.

3.2 RDF-Datentypen

Rdf-Datentypen repräsentieren Werte wie Integer und Fließkommazahlen und bestehen aus drei Bereichen.

Aufbau am Beispiel:

Value Space	{T, F}	Werte
Lexical Space	{"0", "1", "true", "false"}	Beschreibungen /Alias
Lexical-to-Value Mapping	{<"true", T>, <"1", T>...}	Verknüpfung von Wert und Alias

Datentypen werden über der eines Literals angefügten URI-Referenz referenziert. Dabei dient das Literal als Träger der nötigen Informationen.

3.3 C#

Zur Nutzung mit Sharepoint:

> Die Entwicklung in Sharepoint erfolgt komplett mit ASP.NET (Active Server Pages .NET), welches eine serverseitige Technik zum Erstellen von Webanwendungen und –Services darstellt. Dieses Framework basiert auf .NET und unterstützt somit C# als Programmiersprache.

Zur Anbindung an das FOX Framework:

> Das FOX Framework an sich ist in JAVA geschrieben, aber die Schnittstelle erfolgt über eine Client-Server-Architektur. Dabei erfolgt die eigentliche Kommunikation über Standard-HTTP-Abfragen. Dadurch ist die Anbindung an das .NET-Framework, im speziellen an C# ohne Probleme möglich.

Zur Speicherung in RDF:

Dazu muss zunächst geklärt werden, in welcher Form die Informationen in RDF abzulegen sind (XML oder N3?). Des Weiteren fehlen noch Spezifikationen zum Backend, zum Beispiel welche Datenbank zur Verfügung steht. Erst wenn diese Fragen geklärt sind, kann entschieden werden, ob die RDF-Implementierung selbst erfolgen soll oder ob zum Beispiel ein Open-Source-Framework wie „DotNetRDF“ genutzt wird. Dabei spielt eine besondere Rolle in welcher Form die Abfragen erfolgen, also ob dazu die Abfragesprache SPARQL genutzt wird oder aber die semantischen Zusammenhänge gar nicht so komplex sind, dass sich diese Unterstützung lohnt. Denn bei der Nutzung eines externen RDF-Frameworks ist der Konfigurationsaufwand üblicherweise Recht hoch.

3.4 Semantic Web

Ist die maschinenlesbare Auszeichnung von Web-Inhalten. Web-Seiten können also sowohl von Menschen, als auch von Computern gelesen und verstanden werden. Die Bedeutung der Inhalte einer Website wird damit eindeutig festgelegt.

Semantisches Tagging

Linguistisch: Markieren von Worten in einem Text entsprechend ihren Eigenschaften (Subjekt, Prädikat, Objekt) Vor dem aktuellen Hintergrund: Kombination von Social Tagging und Semantic Web: Texte werden automatisiert ausgewertet und verschlagwortet um Wissen zu generieren indem Information verwaltet wird.

4. Aspekte

4.1 Darstellung der Tags (Metadaten)

MS Sharepoint hat eine GUI zum Editieren und Pflegen der Metadaten (Taxonomie), die einfaches Hinzufügen und Ändern von Terms in einer Ordnerstruktur ermöglicht. Die extrahierten Tags könnten also als Terms gruppiert in einer Term-Gruppe (oder mehreren Term-Gruppen) über die Sharepoint-GUI zugänglich gemacht werden, sodass auf die generierten Metadaten übersichtlich zugegriffen werden kann. Terms können über das Sharepoint-Framework hinzugefügt werden.

4.2 Verfügbare Knowledge Extraction Tools

Um die RDF-Tripel aus dem Dokument-Text zu generieren soll ein Framework verwendet werden. Einige Knowledge Extraction Frameworks:

- PPX (PoolParty Extractor): Analyse von Text mit Text-Mining-Algorithmen und semiautomatisierte Annotation von Dokumenten. Als Ausgabeformat ist auch RDF verfügbar. Eingabeformate sind u.a. HTML und einfacher Text.
- OpenCalais: Benutzt NLP, machine learning und andere Prozesse um Entities in einem Dokument (HTML, XML, Text) zu finden. Ausgabe der Entities ist auch als RDF-Tripel möglich.
- FOX: Das Federated Knowledge Extraction Framework ist ein Java-Framework das dieselben Funktionen hat wie die beiden obigen Tools, dh. Keyword Extraction aus Text/HTML durch NLP-Algorithmen und Generierung von RDF-Tripeln.

4.3 FOX

FOX (Federated Knowledge Extraction Framework) ist ein Java-Framework, das mit verschiedenen NLP-Algorithmen aus NL (Natural Language) Keywords extrahiert und darüber hinaus RDF-Tripel generiert. FOX soll für dieses Projekt verwendet werden.

4.4 Programmiersprachen

Plug-ins für Microsoft Sharepoint können in .NET, Silverlight oder Java erstellt werden. Es stellt sich die Frage ob .NET C# genutzt wird, da das Team erfahren in der Nutzung dieser Sprache ist, oder aber Java, da das präferierte Framework FOX auf Java basiert. Microsoft Silverlight kommt für dieses Projekt nicht infrage.

5. Quellen

http://nlp2rdf.org/demo-development

http://wiki.nlp2rdf.org/wiki/DBpedia_Spotlight

http://www.w3.org/TR/WD-rdf-syntax-971002/

http://www.microsoft.com/visualstudio/deu/products/visual-studio-express-products

http://www.csharpme.de/index.php

http://openbook.galileocomputing.de/csharp/

http://msdn.microsoft.com/de-de/library/vstudio/ee231568.aspx

http://msdn.microsoft.com/de-de/library/vstudio/ee330921.aspx

http://139.18.2.164:4444/demo/index.html (Dokumentation von FOX)

http://onezserver.appspot.com/blog.aksw.org/category/projects/fox/

http://www.codeproject.com/Articles/44219/Step-by-Step-SharePoint-Server-2010-Installation-G

Böhnstedt, D. : Semantisches Tagging zur Verwaltung von webbasierten Lernressourcen - Modelle,

Methoden und eine Plattform zur Unterstützung Ressourcen-basierten Lernens. Darmstadt [Dissertation],

2011. URL: http://tuprints.ulb.tu-darmstadt.de/2729/4/2011-Boehnstedt-Dissertation-Semantisches_Tagging_zur_Verwaltung_von_webbasierten_Lernressourcen.pdf (vom 23.12.2012).

Ngonga Ngomo, A.-C., Heino, N., Speck, R., Hillner, S.: Federated Knowledge Extraction for Semantic Web Applications,

http://www.opencalais.com/about

http://en.wikipedia.org/wiki/Knowledge_extraction#Traditional_Information_Extraction_.28IE.29