Benutzerprofile für Semantic Web Anwendungen

Ein Überblick

Bianca Gotthart
Fachhochschule Hagenberg
Softwarepark 11
4232 Hagenberg, Österreich
bianca.gotthart@fh-hagenberg.at

ABSTRACT

Im Web gibt es eine große Menge an Daten. Um für den Benutzer einen Überblick zu ermöglichen, sind Benutzerprofile im Web von besonderer Bedeutung geworden. Mit Hilfe von semantischen Technologien und Konzepten kann auf kontextbezogene und individuelle Interessen vom Benutzer zugegriffen werden, um dies als Basis für die Personalisierung im Web zu verwenden. Diese Arbeit soll einen Überblick geben, welche Arten von Konzepte für die Erstellung von Benutzerprofilen angewendet werden.

Keywords

Web Personalisierung, Benutzerprofile, Semantic Web

1. EINLEITUNG

Eine große Menge an Daten werden täglich im Web veröffentlicht. Sich einen Überblick in der Menge an Informationen zu beschaffen ist für einen Benutzer nur schwer möglich. Zu relevanten Informationen zu kommen ist oftmals nicht einfach. [10]

Personalisierung im Web ist eine Eigenschaft, die im Word Wide Web benötigt wird, um den Informationsüberfluss für den Benutzer zu regulieren. Wenn Wissen vom Benutzer vorhanden ist, ist es möglich, dies als Basis für die Filterung von Informationen zu verwenden. Mit den Technologien des Semantic Webs ist es möglich, nicht nur rein auf die vorhandenen Wörter zu achten, sondern auch den Kontext eines Textes zu ermitteln und somit Verknüpfungen zueinander zu schaffen.

Ein wichtiger Aspekt von Personalisierung ist die Erstellung von Benutzerprofilen. In der Domäne Web bedeutet dies somit das Sammeln von Informationen über einen Benutzer. Das Ziel dieses Vorgangs ist, die Interessen des Benutzers zu erkennen und dies an die Filterung und Darstellung von Webseiten individuell anzupassen. [15]

In dieser Arbeit möchte ich einen Überblick geben, wie die Speicherung von Wissen über Webseiten und Interessen von Benutzer mit Hilfe von Semantic Web Technologien in anderen Projekten umgesetzt wurden und aktuelle Standards erläutern. Der Fokus dieser Arbeit liegt nicht auf die Extrahierung von Daten, sondern vielmehr auf die möglichen Konzepte, wie Benutzerprofile erstellt werden können.

2. SAMMLUNG VON BENUTZERDATEN

Für das Modellieren von Benutzerprofilen müssen Daten über den Benutzer gesammelt werden, sowohl demographische, als auch Wissen über Interessen, Verhalten und Vorlieben, um Informationen personalisiert zu filtern. Das Sammlung von Informationen kann entweder impliizt oder explizit geschehen.

Bei der expliziten Sammlung von Informationen wird der Benutzer gebeten, selbst Interessen zu definieren, indem diese entweder vom Benutzer selbst definiert werden oder diese mit Hilfe von Bewertung einem System mitgeteilt werden. Um für letzteres ein Beispiel zu nennen [15], könnte der Benutzer nach dem Lesen eines Artikels abstimmen, ob dieser die Interessen getroffen hat oder eben nicht. Ein System kann somit dieses Wissen über den Benutzer speichern und auswerten, um bei Suchanfragen die Ergebnisse dahingegen anzupassen. Die Nachteile bei der expliziten Angabe von Interessen ist allerdings die Zeitkomponente. Interessen können sich sehr schnell ändern und ein Profil ständig anzupassen würde für den Benutzer Aufwand und Konsequenz abverlangen.

Im Gegensatz dazu werden bei der impliziten Benutzermodellierung die Interessen mit diversen Algorithmen automatisch analysiert, wobei der Benutzer keinen Eigenaufwand haben soll und dies im besten Falle gar nicht bemerkt. [15] Die Daten werden während des Surfens gesammelt, indem von den Seiten, die der Benutzer betrachtet, Informationen wie zum Beispiel relevante Begriffe, Entitäten, Kategorien gespeichert werden. Deswegen konzentrieren sich viele Recherchen auf die implizite Generierung von Benutzerdaten, da diese die Interessen des Benutzers am deutlichsten widerspiegeln, im besonderen mit semantischen Analysen im Hintergrund, und am wenigsten Zusatzaufwand für diesen bringen.

Bei [6] Ansatz wurde bereits ein adaptiver personalisierter Web Browser entwickelt, der rein auf implizite Weise die Informationen des Benutzers sammelt. Das Benutzerverhalten wird als Benutzerprofil gespeichert. Es wurde ein Messwert entwickelt, der die Interessantheit besuchter Webseiten bestimmt. Sie definieren ein Benutzerprofil mit zwei Komponenten: *Page Interest Estimator* (PIE) und einem *Web Access Graph* (WAG). Basierend auf den PIE, der vom Verhalten des Benutzers erlernt wird, werden die Interessen charakterisiert. Der WAG beinhaltet die besuchten Webseiten mit den gesammelten Mustern und mit Hilfe des PIE

wird deren Relevanz errechnet. Um die Interessen zu sammeln, wurden folgende Quellen berücksichtigt: Browsing History (Worte mit hoher Häufigkeit wurden extrahiert), Seiten die als Lesezeichen abgespeichert wurden, besuchte verlinkte Seiten auf der Webseite und Server Log Dateien (Größe der Seite, Zeitpunkt für die Abfrage der Aktualität, IP Adresse, URL).

Einige Arbeiten erstellen Benutzerprofile als hierarchische Baumstruktur, worin die Interessen des Benutzers gespeichert werden. Kim und Chen [9] unterscheiden allgemeine Interessensgebiete, die langfristig zutreffen und spezifischen Interessensgebieten, die eher von kurzfristiger Dauer sind. Bei diesem Ansatz werden nicht Dokumente geclustert, sondern vielmehr Eigenschaften der Dokumente, worin Dokumente eingegliedert werden. Als Hauptzweige werden die allgemeinen Interessen (langfristigen) und die Unterknoten als spezifische (kurzfristigen) Interessensbegriffe bezeichnet. Dieser Ansatz wird laut Kim und Chen *User Interest Hierarchy* (UIH) genannt. Als Daten werden dabei Webseiten verwendet, die der Benutzer besucht hat und häufig vorkommende Wörter werden extrahiert, die anschließend mit Ähnlichkeitsalgorithmen geclustert in die Hierarchie eingefügt werden. Dadurch kann Zugehörigkeit zwischen Begriffen zu Themen ermöglicht werden.

3. SEMANTIC WEB STANDARDS

Hit Hilfe von Semantic Web Technologien können Benutzerprofile angereichert werden, um kontextbezogene Informationen aus Web-Inhalten zu extrahieren und diese dem Benutzer zuzuweisen. Im folgenden werden Semantic Web Technologien erläutert, die in der Praxis eingesetzt werden.

3.1 Ressource Description Format (RDF)

Ressource Description Format (RDF) [2] ist eine W3C Spezifikation. RDF ist eine formale Sprache, um allgemeine webbasierte Informationen auszuzeichnen. Eine Information wird in einem Triple repräsentieren (Subjekt, Prädikat und Objekt). Das Subjekt bezeichnet eine Ressource im Web mit einer URI und das Prädikat beschreibt die Beziehung zwischen Subjekt und Objekt. Ein Objekt kann wiederum eine Ressource sein oder ist als Literal ausgezeichnet. Im Allgemeinen ist es eine standardisierte semantische Annotationsmethode, wodurch Austauschbarkeit von Informationen ermöglicht wird.

3.2 Ontologie und Web Ontology Language

Eine Ontologie ist eine Menge an Konzepten um Inhalte zu organisieren und zu klassifizieren. Es modelliert Entitäten, logische Verknüpfungen und Beziehungen in Form eines Graphes. Der Standard von *Web Ontology Language* (OWL) [14] im Gegenzug zu RDF ermöglicht es, mehr Bedeutung den Ressourcen zuzuordnen. Eigenschaften können nicht mehr nur auf einzelne Domäne fixiert werden, sondern vielmehr auch das Verhalten zueinander beschreiben. [3]

Um Innovation in einer Blog Community festzustellen [12], ist OWL für die Beschreibung einer Template Ontologie eingesetzt worden. Dafür ist eine Ontologie für die spezielle Domäne entwickelt worden. Dafür ist Expertenwissen notwendig, um alle möglichen Aspekte (Klassen und Eigenschaften) dieser Domäne abzudecken. Die Ontologie wurde verwendet, um eine hierarchische Struktur für die Beziehung zwischen Klassen und Eigenschaften zu erstellen. Die Interessen wurden den Blogartikel von den Benutzern entnommen, indem alle Beiträge klassifiziert und der Ontologie zugewiesen wurden, ohne einer expliziten Einwirkung des Benutzers. Mit Hilfe von Ähnlichkeitsmessungen und Häufigkeitsvorkommnissen einzelner Instanzen in Blogartikeln, wurden innovative Themen herauskristallisiert, welche den Interessen der Community entsprechen.

3.3 Linked Data

Bei dem Konzept von Linked Data handelt es sich darum, Informationen im Web in strukturierter Art und Weise anzubieten. Die Daten sollen von Maschinen lesbar gemacht sein, damit die Bedeutung von Ressourcen auch für Maschinen interpretierbar ist. Des weiteren ermöglicht es eine Verlinkung mit externen Datenquellen. Der Zugriff ist nach Berners-Lee Definition mit einer URI (Unified Ressource Identifier) anzubieten, damit Leute im Web sich die Ressourcen auch ansehen und nützliche Informationen verlinken können.

3.3.1 DMOZ

DMOZ ist ein Open Source Projekt, dass über 4 Millionen Webseiten in 590.000 Kategorien eingeordnet hat. Die Datenstruktur ist als Baum aufgebaut, worin die Kategorien die Knoten und die Webseiten die Blätter sind. Knoten können auf Grund von der reinen symbolischen Verlinkung zu mehreren Elternknoten gehören. [1]

DMOZ ist bereits in einigen Projekt angewendet worden, um eine personalisierte Suche zu ermöglichen. Wie in [7]) angewandt, wodurch der Benutzer selbst Interessen definieren muss, die aus der DMOZ Datenbank aufgelistet werden. Diese Themen werden dann in den jeweiligen Suchalgorithmen miteinberechnet, wie zum Beispiel für die Berechnung des PageRanks bei Google. Als Basis werden verschiedene Algorithmen angewandt, die zum Beispiel auf die Messung der Distanz zwischen zwei Knoten im DMOZ System beruht, je weiter auseinander zwei Knoten liegen, desto weniger haben sie miteinander zu tun.

3.3.2 DBpedia

DBpedia veröffentlich die Daten aus Wikipedia in einer strukturierten und geordneten Art und Weise. Bereits viele Projekte wurden mit dieser Basis entwickelt, um Benutzerprofile oder Web-Ressourcen zu klassifizieren oder anzureichern. Wikipedia hat eine große Menge an Informationen, die mit Kategorien versehen sind. In einem Kategorie System vom Wikipedia sind die Dokumente in zwölf Hauptkategorien unterteilt: Referenz, Kultur, Geografie, Gesundheit, Geschichte, Mathematik, Natur, Menschen, Philosophie, Religion, Gesell- schaft, Technologie. [11] Bei [13] wird ein hierarchisches Benutzerprofil mit Hilfe des Wikipedia Konzeptes erstellt, wobei zuerst Webseiten und Dokumente dem Wikipedia Konzept zugeordnet werden.

3.3.3 Semantic APIs

Um aus einem Text semantische Information herauszubekommen, gibt es bereits eine Vielzahl an Webservices, die dabei Unterstützung bieten, um Entitätenextraktion, Zuweisung von passenden Kategorien oder Empfehlungen für verwandte Artikel im Web zu erhalten. Diese Services basieren auf den

Linked Open Data Systemen oder haben ihre eigene Ontologie aufgebaut. Im Allgemeinen beiten diese Webservices eine semantische Analyse von Texte bzw. URLs im Web an, die folgende Arten als Output generieren: Textextraktionen (wie zum Beispiel Entitäten-Extraktion, Schlüsselwörter- Extration) oder Textklassifikationen in unterschiedlicher Form (RDF, JSON, XML). [8]

Die bekanntesten APIs sind OpenCalais [1], Zemanta [2], Alchemy API [3] und OpenAmplify [4]. OpenCalais bietet ebenfalls eine eigene Ontologie an[5].

3.3.4 FOAF Ontologie

Das Projekt *TUMS*, wie in [17] beschrieben, bietet ein Service an, dass mit Hilfe von Twitter Nachrichten ein Benutzerprofil erstellt, indem die Interessen auf Basis einer Ontologie zugewiesen werden. Für die Repräsentation des Benutzers wird auf das Friend-Of-A-Friend (FOAF) Projekt [4] zurückgegriffen, die Informationen über den Benutzer zu sammeln. Dieses Konzept beinhaltet sowohl allgemeine Informationen, aber auch eine Möglichkeit, die Interessen mit Hilfe des *Weighted Interest Vocabulary* [5] zu gewichten. Die Informationen werden in diesem Ansatz entweder themenbasiert, tagbasiert oder entitätenbasiert gespeichert und als Gewichtung wird die Begriffshäufigkeit, die ein Benutzer verwendet, angewandt. So könnte eine mögliche Auszeichnung eines Benutzers aussehen (entnommen von [17]):

```
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
    @prefix wi: <http://purl.org/ontology/wi/core#> .
    @prefix wo: <http://purl.org/ontology/wo/core#> .
    @prefix ex: <http://example.org/> .
    @prefix dbpedia: <http://dbpedia.org/resource/>
    <http://twitter.com/taubau>
      a foaf:Person ;
      foaf:name "Ke Tao";
      wi:preference [
         a wi:WeightedInterest ;
         wi:topic dbpedia:Jazz ;
         wo:weight [
a wo:Weight ; wo:weight_value 10.0 ; wo:scale ex:AScale
].
```

Oftmals werden Ansätze in gemischter Form umgesetzt. Wie bei [16] liegt der Fokus darin, dass eine Architektur mit Hilfe von den Benutzerinteraktion in den unterschiedlichen Folkonomy Seiten erstellt wird. Um die tatsächlichen Interessen eines Benutzers zu erhalten, werden die Tags aus diesen Seiten entnommen und einer semantischen Analyse und Zuordnung unterzogen. Um eine allgemeine Lösung für die Zuordnung von Tags und deren Beziehungen zu schaffen, wird dafür auf die Wikipedia Kategorien zurückgegriffen, da Wikipedia eine große Palette an Themen anbietet. Ein wichtiger Bestandteil dieses Ansatzes ist, dass Tags im vorhinein gefiltert und zusammengefügt werden und auf Grund von Ähnlichkeiten zwischen Begriffen wie zum Beispiel *secondlife* und *second-life* vom System als das Gleiche angesehen werden. Mit Hilfe der FOAF-Ontologie wird das neue Thema dem Benutzerprofil gespeichert, wofür eine RDF Profil mit dem FOAF-Attribut *interest* verwendet wird, um die Wikipeda Kategorie mit dem Benutzer zu verknüpfen. Eine Kategorie erhält noch eine Gewichtung, je nachdem wie oft ein Tag dieser Kategorie hinzugefügt und dieser vom Benutzer verwendet wurde.

[1]http://www.opencalais.com/
[2]http://www.zemanta.com/
[3]http://www.alchemyapi.com/
[4]http://www.openamplify.com/
[5]http://www.opencalais.com/documentation/opencalais-web-service-api/calais-ontology-owl

4. CONCLUSION

In diesem Paper wurden mögliche Konzepte, wie Benutzerprofile aufgebaut werden können, erläutert. Der Fokus lag auf allgemeinen Informationen über die Sammlung von Daten, bis hin zu semantischen Technologien, die verwendet werden können, um kontextuell Informationen zu erhalten und zu speichern. Mein Ansatz wird mehrere Theorien vermischen. Mit Hilfe von ausgewählten Semantic APIs wird Textextraktion durchgeführt und mit Hilfe des Wikipedia Konzeptes kategorisiert. Als Input Daten wird zum einen die Browser History des Benutzers angesehen, um die Interessen festzustellen, und zum anderen werden RSS-Feeds kategorisiert, um diese für den Benutzer als Information täglich anzubieten. Als Technologie wird eine minimalistische Ontologie (mit Entitäten, Themen, Schlüsselwörter) verwendet, worin sowohl das Benutzerprofil, als auch die RSS-Daten gespeichert werden, um eine Verbindung zwischen diesen zwei Konzepten zu generieren.

5. REFERENZEN

[1] Open directory project. http://dmoz.org/, 2013.

[2] D. Backett and B. McBride. RDF/XML syntax specification. www.w3.org/TR/rdf-syntax-grammar/, 2004.

[3] T. Berners-Lee. Linked data. http://www.w3.org/DesignIssues/LinkedData.html, 2006.

[4] D. Brickley and L. Miller. The friend of a friend (FOAF) vocabulary specification. http://xmlns.com/foaf/spec/, Aug. 2010.

[5] D. Brickley, L. Miller, T. Inkster, Y. Zeng, Y. Wang, D. Damljanovic, Z. Huang, S. Kinsella, J. Breslin, and B. Ferris. The weighted interests vocabulary 0.5. http://smiy.sourceforge.net/wi/spec/weightedinterests.html, Sept. 2010.

[6] P. K. Chan. *A Non-Invasive Learning Approach to Building Web User Profiles*. 1999.

[7] P. A. Chirita, W. Nejdl, R. Paiu, and C. KohlschÃijtter. Using ODP metadata to personalize search. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '05, page 178âĂŞ185, New York, NY, USA, 2005. ACM.

[8] F. Dotsika. Semantic APIs: scaling up towards the semantic web. *Int. J. Inf. Manag.*, 30(4):335âĂŞ342, Aug. 2010.

[9] H. R. Kim and P. K. Chan. Learning implicit user interest hierarchy for context in personalization. In *Proceedings of the 8th international conference on Intelligent user interfaces*, IUI '03, page 101âĂŞ108, New York, NY, USA, 2003. ACM.

[10] M. Marchiori. Towards a people's web: Metalog. In *Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence*, WI '04, page 320âĂŞ326, Washington, DC, USA, 2004. IEEE Computer Society.

[11] J. Min and G. J. Jones. Building user interest profiles from wikipedia clusters. July 2011.

[12] M. Nakatsuji, M. Yoshida, and T. Ishida. Detecting innovative topics based on user-interest ontology. *Web Semant.*, 7(2):107âĂŞ120, Apr. 2009.

[13] K. Ramanathan and K. Kapoor. Creating user profiles using wikipedia. In *Proceedings of the 28th International Conference on Conceptual Modeling*, ER '09, page 415âĂŞ427, Berlin, Heidelberg, 2009. Springer-Verlag.

[14] M. K. Smith, C. Welty, and D. L. McGuinnes. OWL web ontology language guide. http://www.w3.org/TR/owl-guide/, 2004.

[15] C. Srinvas. Explicit user profiles for semantic web search using XML. *International Journal of Engineering Research and Applications (IJERA)*, 2(6):234–241, Dec. 2012.

[16] M. Szomszor, H. Alani, I. Cantador, K. O'Hara, and N. Shadbolt. Semantic modelling of user interests based on cross-folksonomy analysis. In *Proceedings of the 7th International Conference on The Semantic Web*, ISWC '08, page 632âĂŞ648, Berlin, Heidelberg, 2008. Springer-Verlag.

[17] K. Tao, F. Abel, Q. Gao, and G.-J. Houben. TUMS: twitter-based user modeling service. In *Proceedings of the 8th international conference on The Semantic Web*, ESWC'11, page 269âĂŞ283, Berlin, Heidelberg, 2012. Springer-Verlag.