Semiautomatische Erweiterung von Topic Maps mit Hilfe von Thesauri und User-Feedback

Andreas Bertram

andy@asware.net

Abstract: Das Semantic Web stellt die nächste Stufe der Evolution des World Wide Web dar. Der Prozess der Erweiterung bestehender Datenbestände um Metadaten - die Annotation - ist teuer und aufwändig. Diese Arbeit untersucht einen Ansatz, in dem ein initiales, von Experten gefertigtes semantisches Netz konventionelle Suchvorgänge begleitet und unterstützt. Das semantische Netz wächst dabei mit seiner Benutzung, der Aufwand bei der Erfassung von Metainformationen wird dadurch reduziert. Die Nutzung des Netzes besteht in den Anwendungsfällen Navigation und Suche im konventionellen Datenbestand. Essentiell für den Erweiterungsprozess ist User-Feedback: Nutzer werden bei der Ergänzung des Netzes durch einen Thesaurus unterstützt. Eine prototypische Implementierung verwendet Topic Maps zur Modellierung des Netzes und stützt sich bei der Suche auf Google. Suchoperationen ergänzen das semantische Netz um Occurrences (Dokumente), während Navigationsoperationen die Erweiterung um Associations (Verknüpfungen) zur Folge haben. Alle erweiterten Objekte unterliegen User-Feedback zur Gewichtung und Qualifizierung. Unterstützend wird der Thesaurus WordNet eingesetzt. Abschließend werden Ergebnisse der Evaluierung vorgestellt.

1 Ziel der Arbeit

Das Internet hat sich in den letzten Jahren, insbesondere mit Hilfe des World Wide Web (WWW)), als preiswerte und für jedermann zugängliche, umfangreiche Informationssammlung etabliert. Die einfach zu erlernende und anzuwendende Dokumentenbeschreibungssprache HTML ermöglichte es, schnell und effizient Inhalte einem breitem, weltweitem Publikum zur Verfügung zu stellen. Das einfache und gleichermaßen mächtige Konzept der Hyperlinks stellt dabei die Navigationsmöglichkeiten zur Verfügung und hat somit entscheidend zur ungebrochenen Popularität und zu einem starken Wachstum des World Wide Web beigetragen. Essentiell für die Evolution vom World Wide Web zum Semantic Web ist die Annotation. Dieser erweitert Dokumente um Metadaten und belegt erweiterte Hyperlinks mit semantischem Mehrwert. Das größte Problem bei diesem Evolutionsschritt ist der Aufwand für die Erfassung und Pflege der Metainformationen. Die Grundidee dieser Arbeit ist, ein semantisches Netz auf einer begrenzten Domäne suchbegleitend und dabei gleichermaßen die Suche unterstützend halbautomatisch aufzubauen. Der Aufbau dieses Netzes startet dabei nicht bei Null, sondern auf einem Grundgerüst semantischer Informationen - einem vorgegebenen, von Experten manuell erzeugten semantischen Netz. Bei diesem Prozess wird eine konventionelle Suchmaschine, ein Thesaurus sowie User-Feedback eingesetzt. Die Kernpunkte dieser Arbeit sind: (1) Gegeben: Semantisches Startnetz auf Datenbestand,

(2) Suche auf semantischem Netz + Suche auf Daten, (3) Vorschlagsgenerator nutzt Thesaurus für neue Themen und (4) User-Feedback zur Erweiterung des semantischen Netzes.

2 Ansatz

Zur Modellierung des semantischen Netzes werden Topic Maps eingesetzt. Die Standardoperationen Navigation und Suche werden funktionell erweitert und um eine dritte Komponente, die Bewertung, ergänzt. Nach Studien u.a. der GVU[1] sind Suchmaschinen und Links in Webseiten die wichtigsten Mittel, um Webseiten zu erreichen. Da bietet es sich an, die Operationen Navigation und Suche funktional zu erweitern, um daraus bei der Erweiterung eines unterliegenden semantischen Netzes zu profitieren. Die funktionale Erweiterung liegt dabei in der Nutzung des semantischen Netzes selbst. Die Bewertung von Dokumenten und Verknüpfungen, welche auch maschinell unterstützt wird, dient der Etablierung und Löschung von Kandidaten und damit der Erweiterung des Netzes. Das System verwaltet zwei Komponenten:

- den konventionellen Datenbestand
- ein semantisches Netz für diesen Datenbestand, welches zusammengesetzt ist aus (1) dem etablierten semantischen Netz, von Expertenhand bestätigt (=Startnetz), und (2) dem potentiellen erweiterten Netz, dessen Knoten und Kanten durch Navigations- und Suchoperationen der vorherigen Nutzer als Kandidaten für die Etablierung eingefügt wurden.

Ein von Expertenhand gefertigtes Startnetz bildet die Grundlage für alle weiteren Operationen und erweitert den zugrundeliegenden Datenbestand um Metainformationen, indem Dokumente zu Themen zugeordnet werden und des weiteren Abhängigkeiten zwischen Themen und damit auch zwischen den Dokumenten formuliert werden. Knoten und Kanten, welche im weiteren Prozess im Netz ergänzt werden, erhalten Gewichte, welche dem essentiellen User-Feedback ausgesetzt sind.

2.1 Navigation

Die Mechanismen, die zur Erweiterung des semantischen Netzes durch Navigation führen, sind in Abbildung 1 dargestellt. Ausgangspunkt ist *Dokument 1*. Dieses ist optimalerweise bereits im semantischen Netz erfasst, d.h. es existiert mindestens ein Topic, das als Occurrence die Adresse von *Dokument 1* enthält. Falls dies nicht der Fall sein sollte, kann es von Hand in das Netz eingefügt werden: es wird ein neues Topic erzeugt und benannt. Dieses erhält als Occurrence das aktuelle Dokument. Der Benutzer navigiert nun unter Nutzung eines im Ausgangsdokument vorhandenen Links zu *Dokument 2*. Für dieses gilt, wie auch für *Dokument 1*, dass es manuell im semantischen

[1] http://www.gvu.gatech.edu/user_surveys/User_Survey_Home.html

Netz erfasst werden kann, falls es dort noch nicht hinterlegt ist. Die Handlung der Navigation von *Dokument 1* zu *Dokument 2* impliziert eine semantische Verbindung zwischen beiden Dokumenten, wenn auch unbekannter Qualität. Sofern noch keine Kante zwischen den beiden Knoten, die den jeweiligen Dokumenten zugeordnet sind, existiert, wird diese nun vom System eingefügt. Die Association, die zur Modellierung dieser neuen Kante in das semantische Netz eingefügt wird, enthält die beiden Dokumenten-Topics, welche als Member in ihren jeweiligen Rollen *als navigation-source* respektive als *navigation-destination* teilnehmen. Die neue Association wird zunächst generisch typisiert unter dem Topic *navigation-related* und mit einem Initialgewicht versehen.

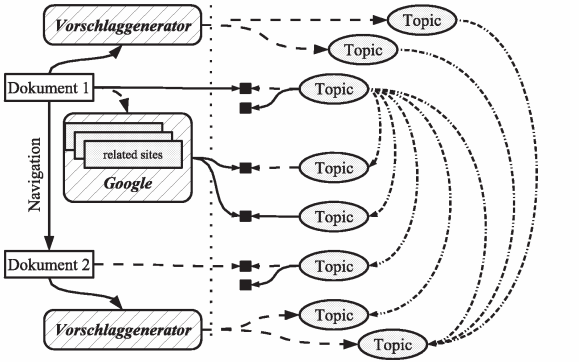

Abbildung 1 Erweiterung durch Navigation

2.2 Suche

Die Suche im semantischen Netz erfolgt über ein Matching der Suchworte zu den Namen der Knoten im Netz. Im Erfolgsfall führt dies zu *semantischen Treffern* und liefert Konzepte, welche auf Dokumente verweisen, und/oder mit anderen, thematisch nahen Konzepten verknüpft sind. Die Suche im Datenbestand erfolgt mit konventionellen Mitteln, einer Volltextsuchmaschine. Falls auf diese Weise Dokumente gefunden werden, welche im semantischen Netz bereits enthalten sind, so erhalten diese den höchsten Stellenwert in der Liste der Ergebnisse. Wenn die Dokumente nicht im semantischen Netz enthalten sind, setzt erneut der Vorschlagsgenerator ein und bietet Konzepte an, mit denen die Dokumente verknüpft und damit annotiert werden können. Während die Navigation das System veranlasst, neue Kanten in das semantische Netz aufzunehmen und Bezüge zwischen Konzepten herzustellen (Associations in der Topic Map), dient die Suche primär der Aufnahme von Ressourcen, also von Dokumenten (Occurrences), respektive der Zuordnung von Dokumenten zu Konzepten. Eine Übersicht über die Wirkung von Suchoperationen zeigt Abbildung 2. Zunächst wird der Anfrageterm selbst als eigenes, typisiertes *queryphrase-topic* in der Topic Map angelegt - falls es dort noch nicht existiert. Wenn das Topic bereits existiert, wurde die Suche schon zuvor ausgeführt und unterliegt bereits Feedback-Mechanismen, welche dafür sorgen sollen, dass die Qualität der verlinkten Topics und Occurrences steigt. Die Anfrage wird an Google weitergeleitet. Die Suchmaschine bietet als Ergebnis

Dokumente an, welche in Form gewichteter Occurrences direkt dem Anfrageterm und damit dem Anfrage-Topic zugeordnet werden. Falls es bereits bestand und nicht erst mit dieser Anfrage erzeugt wurde, so werden die bestehenden Occurences ergänzt, sofern diese noch nicht dem Topic zugeordnet sind. Diese eventuelle Ergänzung wird durchgeführt, da sich Suchergebnisse zu einer Anfrage aufgrund der Dynamik des World Wide Web im Laufe der Zeit ändern können, und sich dies auch im semantischen Netz wiederfinden soll. Basierend auf den von Google gelieferten Dokumenten wird eine Suche im semantischen Netz durchgeführt, um nach den Topics zu suchen, denen ein Dokument des Ergebnisraums zugeordnet wurde. Dies geschieht durch ein Matching der Occurrences des Anfrage-Topics mit den restliches Topics des semantischen Netzes. Im Falle eines Treffers kann man auf eine semantische Verbindung - zunächst unbekannter Art - zwischen dem Anfrage-Topic und dem gefundenen Topic schließen. Viele gemeinsame Occurrences von Anfrage-Topic und in diesem Status gefundenen Topics könnte auf Gleichheit des Konzepts hin deuten - die Qualität der Verbindung ist unbekannt und es gibt keine Möglichkeit, diese algorithmisch festzustellen. Es obliegt daher dem Nutzer, eine entsprechende Verbindung manuell anzugeben und zu qualifizieren. Nun kommt der linguistische Thesaurus ins Spiel. Auf Anforderung hin wird die Suchanfrage mit Hilfe von WordNet in *Indexworte* zerlegt. Für jedes Indexwort wird ein Indexwort-Topic angelegt, welches jeweils mit dem Anfrage-Topic verknüpft wird, durch das es erzeugt wurde. Die hierbei erzeugte Association wird über das Topic *indexword-related* typisiert. Die Verbindung von Anfrage-Topics zu Indexwort-Topics führt möglicherweise zu weiteren Anfrage-Topics, die mit demselben Indexwort-Topics verbunden sind. Die Motivation hierbei ist es, potentiell semantisch verwandte Anfrageterme miteinander zu verknüpfen. Im nächsten Schritt können mit Hilfe des Thesaurus von jedem Indexwort erweiterte Wörter im semantischen Umfeld ermittelt werden. Das System sucht dabei nach Synonymen (andere Wörter gleicher Bedeutung), Hypernymen (Überbegriffen) und Hyponymen (Unterbegriffen). Diese dienen der Generalisierung respektive einer Spezialisierung der Suche. Für diese Worte werden ebenso Topics erzeugt und entsprechende Verknüpfungen mit dem Ausgangs-Indexwort angelegt. Bei diesen neuen, ungewichteten Associations kommen die Topics *wordnet-hypernym*, *wordnet-hyponym* und *wordnet-synonym* zwecks Typisierung der neuen Wort-Topics zum Einsatz. Auf diese Weise wird der WordNet-Thesaurus mit der Zeit komplett in der Topic-Map nachgebildet (gleiches gilt für den zuvor skizzierten Einsatz der dmoz-Informationen).

2.3 Feedback

Der Prozess zur Aufnahme von neuen Themen und Relationen kann nicht vollautomatisch anhand des Algorithmus (Thesaurus als Vorschlagsgenerator) erfolgen, weil ein Thesaurus alleine nur Hinweise, nicht aber vollständige Lösungen für die semantische Verknüpfung geben kann. Die Bewertung der Relevanz sowohl eingefügter Knoten als auch maschinell ergänzter Kanten obliegt also den Nutzern des Systems (Suchende sowie Experten zwecks Wartung). Daher müssen die vorgeschlagenen Konzepte und Relationen durch explizite Rückmeldungen der Nutzer überprüft werden. Dieses *Feedback* ist die menschliche und damit intelligente Komponente im Lernprozess des Systems, welche ein semantisches Netz erweitern und den Nutzer gleichzeitig bei

seiner Navigation und Suche unterstützen soll. Der gesamte Prozess lebt von dieser Interaktion zwischen Mensch und Maschine.

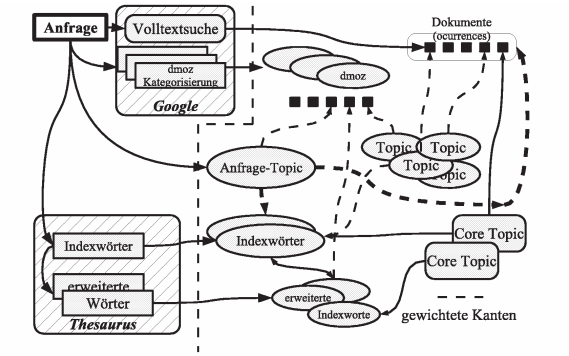

Abbildung 2 Erweiterung durch Suche

3 Prototyp

Zur Evaluierung des Ansatzes wurde ein Prototyp implementiert. Dieser wurde einer Reihe von Testläufen unterzogen, bei denen das Wachstum des semantischen Netzes in Abhängigkeit von Navigations-, Such- und Feedbackoperationen protokolliert wurde. Zu diesem Zweck standen drei semantische Startnetze im XTM-Format zur Verfügung. Es wurden die Auswirkungen der Operationen *Navigation*, *Suche* und *Feedback* und ihre Wirkung auf das semantische Netz hinsichtlich Anzahl Topics, Anzahl Associations, Anzahl Occurrences, Qualität der Associations (Kanten) und Qualität der Occurrences (Knoten, Zuordnung der Dokumente zu Themen) bewertet (siehe Abbildung 3). Die Auswertung ergab, dass der Ansatz ein langsames feedbackgestützter Wachstum des semantischen Netzes unterstützt.

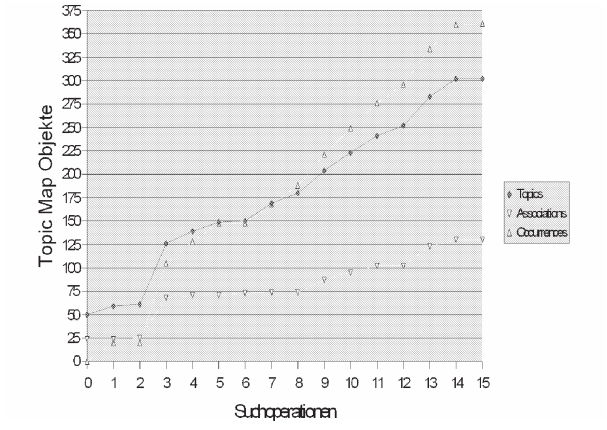

Abbildung 3 Wachstum des Netzes bei Suchoperation