Inhalt

1. Motivation - Vision

Im Internet kommt es zu einer Dateninflation – nie zuvor standen so viele Informationen zur Verfügung und nie war es so schwierig, daraus das wirklich Relevante zu filtern. Neue Technologien, basierend auf aktuellen Forschungen in der Informatik, werden es für die Benutzer künftig einfacher machen, genau die Information zu finden, die sie tatsächlich brauchen.

Wie der Informatik-Visionär Stephen Wolfram in seinem Blog zur kommenden Super-Suchmaschine (wolframalpha) schreibt, ist es dazu notwendig, Methoden und Modelle zu finden um die Billionen vorliegender Textdaten mathematisch verarbeitbar zu machen. Ausgehend von dieser Vision haben wir in unserem Projekt eine mathematische Verarbeitung von Kundenrezensionen von online-shops entwickelt, die es dem Benutzer ermöglicht aus einer Vielzahl rein textueller, qualitativer Aussagen (z.B. „Der Sound ist super“) eine quantitative Bewertung eines Produktmerkmals (Sound: 9/10 Punkten) zu erhalten.

(http://blog.wolfram.com/2009/03/05/wolframalpha-is-coming/)

2. Semantik – Web 3.0

Unter Semantik versteht man die Wissenschaft von der „Bedeutung“ in der Kommunikation, sie beschäftigt sich mit der Bedeutung der Zeichen und Ausdrücke einer Sprache. In der Informationstheorie versteht man darunter die Bedeutung einer Informationsfolge, im Gegensatz zum Informationsgehalt (der auch bei einer Zufallsfolge sehr hoch sein kann, aber keinerlei Bedeutung und damit keine Semantik aufweist)

Unter dem „Semantischen Web“ bzw. dem „Web 3.0“ versteht man eine Erweiterung der bisherigen Strukturen im Internet um semantische Komponenten. Informationen sollen nicht nur von Menschen verstanden sondern auch von Maschinen verarbeitet werden können. Web 3.0 setzt allerdings Annotationen im Text voraus, ist also nicht mit den bisherigen Informationen im Internet kompatibel und daher ist eine Umsetzung in naher Zukunft fraglich. In unserem Projekt werden daher keine semantischen Annotationen vorausgesetzt, sondern wir bedienen und statistischer und semantischer Zusammenhänge in normalem Alltagstext.

http://logik.phl.univie.ac.at/~chris/skriptum/node26.html

http://de.wikipedia.org/wiki/Semantik#Informationstheorie

3. Tag – Text – Cloud

Unter einer Tag (oder Text) Cloud versteht man eine Methode um Informationen zu visualisieren. Eine Liste aus Schlag- bzw. Stichworten wird dabei alphabetisch sortiert, wobei Wörter die häufiger vorkommen größer dargestellt werden.

Tag Clouds werden heute bei vielen Webseiten eingesetzt um Neuigkeiten oder wichtige Informationen herauszustreichen, sie haben aber den Nachteil, dass sie (ziemlich) statisch sind und die Bedeutung zwischen den Schlüsselwörtern in keiner Weise berücksichtigt wird.

4. ATC

Im Projekt „Advanced Tag Cloud“, welches von uns im Jahr 2008 entwickelt wurde, werden Texte einer statistischen Analyse unterzogen, d.h. sie werden auf Wortanzahl, Häufigkeit, und Sinnhaftigkeit überprüft und die Beziehungen, die zwischen zwei Wörter bestehen analysiert.

Das Front-End setzt dann die berechnete Information aus der Datenbank optisch um, sodass der Benutzer nicht nur sieht, wie die Wörter dual in Beziehung stehen, sondern welche Beziehungen sie im Cluster einnehmen. Durch intelligente Auswahl der passenden Wörter werden dem Benutzer nun nur noch diejenigen Textstellen vorgeschlagen, in denen alle Wörter des gewählten Clusters in entscheidender Weise vorkommen.

So können durch Strukturierung und Visualisierung große Mengen unstrukturierter Texte miteinander in Beziehung gesetzt werden; der Benutzer sieht nicht mehr tausende untaugliche Suchergebnisse, sondern nur noch die Treffer, die für ihn wirklich relevant sind.

Wir danken cyberschool.at für die Prämierung des Projektes mit dem ersten Preis der Sparte Technics im Mai 2008 (http://derstandard.at/?url=/?id=1234509443804) .

5. STC - Semantic Tag Cloud

Unsere Semantic Tag Cloud ist eine dynamisch generierte Stichwortwolke (aufbauend auf ATC), die sowohl die Häufigkeit der Wörter als auch deren Vorkommen in den entsprechenden Texten, vor allem aber die semantischen Zusammenhänge untereinander berücksichtigt. Demnach werden nicht nur die Abstände zueinander berechnet, sondern deren tatsächliche Bedeutung in Kombination (Nomen zu Adjektiv) ausgewertet.

Es erfolgt eine statistische Analyse der zu untersuchenden Texte und danach eine aufwendige mathematische Bewertung der Textzusammenhänge, unter Berücksichtigung der deutschen Grammatik. Dazu zählt neben der Nomen – Adjektiv Kombination auch Negationen (nicht, wenig, …) und Verstärkungen (sehr, total, …) der Adjektive. Auch ein Abgleich der Schlüsselwörter mittels Synonymliste findet statt, um eine redundante Auswertung zu verhindern.

Hauptanwendung dieses Projekts ist die (quantitative) Auswertung der (qualitativen) Kundenrezensionen von Online Shop. Dabei werden die Bewertungen mithilfe einer „Anti-Keyword-List“, welche alle irrelevanten Wörter beinhaltet, gefiltert und in der Datenbank gesammelt. Die wichtigen Keywords (z.B. bei einem MP3-Player: Klangqualität, Größe, Preis-/Leistungsverhältnis, usw. …) werden anschließend grafisch dargestellt und im Zusammenhang mit den zugehörigen Bewertungen angezeigt.

6. Input – C# - Datenbank

Um eine Textanalyse der Kundenrezensionen durchzuführen werden zuerst die Texte vom entsprechenden Onlineshop extrahiert und anschließend aus den gespeicherten Texten die notwendigen Metaattribute herausgefiltert - entweder mittels Anti-Keyword-List oder mittels themenspezifischer Keylist – Vorlagen. Ersteres dient zur Erstellung einer neuen Vorlage für eine noch nicht vorhandene Produktsparte und wird vom Administrator durchgeführt, Letzteres dient dem Einsatz beim Enduser.

Die Entwicklungsarbeit findet in Visual Studio C# sowie mit einer Microsoft Office Access Datenbank statt, die diverse relativ statische sowie dynamische Tabellen beinhaltet. Die statischen Tabellen sind bereits vor Anwendung der Software vorhanden und werden zur Textanalyse benötigt. Jede Produktsparte hat eine eigene Vorlage, aus der die relevanten Wörter für die Textanalyse herangezogen werden, parallel dazu existieren Adjektiv-, Negationen- und Verstärkungstabellen.

In der Adjektivtabelle befindet sich eine Vielzahl wichtiger Adjektiven, die bei der semantischen Textauswertung mit den zugehörigen Nomen zur Bewertung verwendet werden. Jedem Adjektiv wurde ein Wert zugewiesen, der die Ausdrucksstärke des Wortes repräsentiert. Demnach ergeben sich trotz positiver Bewertung verschiedene Werte für verschiedene Kombinationen von Nomen und Adjektiven. Außerdem werden Negationen berücksichtigt, steht beispielsweise ein Nomen in Kombination mit einem Adjektiv im Text, wird kontrolliert, ob das Adjektiv in positiver Form gemeint ist, d.h., dass sich kein Wort wie „nicht“ oder „kein“ vor dem Adjektiv steht. Außerdem berücksichtigen wir die Zusammenhänge von Adjektiven und Verstärkungen wie „sehr“. All diese Wörter befinden sich in den statischen Tabellen der Datenbank.

Die dynamischen Tabellen werden im Laufe der Textanalyse gefüllt, also Texttabelle, Keylist, Positionstabelle und Bewertungstabelle. Letztere wird für Auswertung und für die grafische Darstellung verwendet.

7. Textauswertung – Synonyme – Wortarten - Keyword /AK

Für „jede“ Produktsparte wird eine eigene Vorlage mit relevanten Wörtern in der Datenbank angelegt. Diese werden nun mit den Wörtern aus den Rezensionen verglichen. Sollte nun eines dieser Wörter aus der Tabelle im Text vorkommen, so wird dafür ein Eintrag in der „Keylist“ angelegt und gegebenenfalls die Häufigkeit bestimmt. In den entsprechenden Vorlagen wird jedem Metaattribut ein Ober-Synonym (Familie) zugeordnet. Damit wird eine redundante Auswertung für verschiedene Metaattribute mit gleicher Bedeutung unterbunden, auch die unterschiedlichen Schreibweisen der Rezensenten können dadurch ausgeglichen werden, in Zukunft ist es dadurch sogar möglich mehrsprachige Analysen durchzuführen

Parallel zu diesem Vorgang wird ebenfalls mit der „Adjektiv-Liste“, der „Negationen-Liste“ und der „Verstärkungs-Liste“ verglichen. Dabei stehen die einzelnen Wörter für Keyword, Adjektiv, Negation und Verstärkung.

Nach Fertigstellung der Keylist hat der User nun noch die Möglichkeit, für ihn irrelevante Wörter zu entfernen, um so die nicht notwendige Rechenzeit zu verkürzen.

Die nächste Phase ist die Bestimmung der Positionen der Keywords. Dabei wird für jedes Vorkommen ein eigener Eintrag angelegt, in dem sowohl die entsprechende KW-ID (aus der Keylist), die Text-ID, sprich die jeweilige Rezension, in der das Wort erwähnt wird, und die Position im Text gespeichert wird. Anhand dieser Positionen können die Abstände der Wörter untereinander berechnet werden. Nachdem nun alle Positionen der Keywords bekannt sind, können die Abstände der Wörter zueinander berechnet werden.

8. Bewertung

Die Bewertung funktioniert auf der Basis, dass jedes Adjektiv einen Wert von minus zehn bis zehn hat, der dessen Bedeutung in Bezug auf ein zugehöriges Nomen ausdrückt. Zur Bewertung wird nun ein Nomen nach dem anderen ausgewählt, und dieses mit jedem vorhandenen Adjektiv in einer Entfernung von N Wörtern kombiniert (dabei ist lineare, quadratische und Gauß-Glockenkurve Auswahl möglich). Befindet sich außerdem in einem Abstand von N Stellen eine Negation oder eine Verstärkung, so wird dies für das Adjektiv vermerkt. Sollte eine Negation vorhanden sein, so wird der entsprechende Wert des Adjektivs negativ gerechnet, bei einer Verstärkung entsprechend der Wortstärke multipliziert.

Mittels entsprechender Abfrage wird nun für jeden Oberbegriff ein Wert ermittelt und aufgrund dieser Zahl kann dann auf die Qualität der Eigenschaft geschlossen werden und so ein Urteil über das Produkt selbst gefällt werden. Die Summe dieser einzelnen Werte gibt schließlich eine allgemeine Wertung über das Produkt selbst ab.

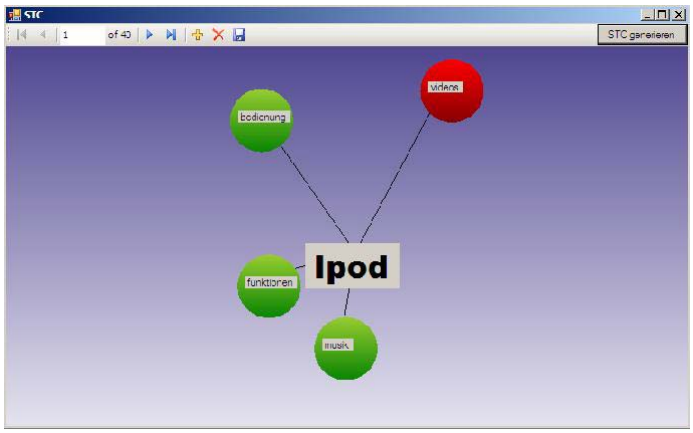
9. Output - Grafische Darstellung

Nachdem nun für jedes im Text vorkommende Nomen, welches in Zusammenhang mit einem Adjektiv auftritt, ein Wert berechnet wurde, können diese nun in einer benutzerfreundlichen und leicht verständlichen Form für die Anwender und potentiellen Käufer an der Applikationsoberfläche dargestellt werden.

Dies geschieht mittels kostenloser Bibliothek zur Erstellung von Diagrammen (ZedGraph). Nach aufwändiger Berechnung werden die Ergebnisse der Metaattribute in einem Balkendiagramm dargestellt, dabei spiegelt die Höhe der Balken die berechneten Werte wieder.

Neben der grafischen Darstellung mittels Balkendiagramm hat der Benutzer die Möglichkeit, die Auswertung in Form der „Semantischen Tag Cloud“ zu betrachten. Dabei

befindet sich das Produkt im Zentrum, während alle Metaattribute mittels Formel rund um das Zentrum im jeweiligen Abstand der Bewertung angeordnet werden. Positive Werte werden durch grüne Kugeln dargestellt, negative durch rote.



Aufgrund dieser grafischen Darstellung kann der potentielle Käufer nun auf die Qualität des Produktes schließen.

10. Kundennutzen

Der Hauptanwendungspunkt des Projekts liegt in der Analyse der Kundenzufriedenheit bei Plattformen von Großhändlern (z.B.: Quelle, Amazon). Produkthersteller oder Händler können mithilfe der Software die Zufriedenheit ihrer Kunden bewerten, indem sie die Rezensionen analysieren, aber im Gegensatz zu bisherigen Methoden nicht basierend auf Sternchen-Bewertung oder langwierigem Durchlesen der Texte, sondern auf einer quantitativen Analyse der Wortzusammenhänge und damit der oft versteckten „soft-fact-information“. Dies hat natürlich einen positiven Einfluss auf das Geschäft, sofern sich Entwicklung & Marketing tatsächlich an den Ergebnissen der Auswertung orientieren.

11. (End-) Benutzernutzen

Aufgrund dieser grafischen Darstellung hat ein potentieller Kunde die Möglichkeit, das für ihn „optimale“ Produkt aufgrund der einzelnen Bewertungen der Produkteigenschaften zu finden und somit aus der Erfahrung vorheriger Käufer zu profitieren da deren textlastige Rezensionen innerhalb kürzester Zeit ausgewertet und in einer für den Benutzer leicht verständlichen Form dargestellt werden.

Der Benutzer braucht also nicht alle der Kundenrezensionen lesen, sondern kann diese einfach mittels STC-SW auswerten lassen, die Software filtert innerhalb kürzester Zeit alle relevanten Produktinformationen und stellt diese übersichtlich an der Oberfläche dar.

Ein weiterer Vorteil ist, dass unsere Software nicht nur eine Gesamtanalyse und somit ein Ergebnis für das Produkt selbst ermittelt, sondern dessen Eigenschaften einzeln auch bewertet. Der Benutzer kann nun auf einem Blick sehen, ob eine Eigenschaft positiv bewertet wurde (positive Werte werden durch grüne Kugeln, negative durch rote Kugeln dargestellt) und durch den Abstand zum Zentrum ist außerdem ersichtlich, ob eine Eigenschaft durchschnittlich oder deutlich positiv/negativ bewertet wurde.

12. Fun

Einarbeitung, Ausarbeitung, Umsetzung und Berichterstellung dieses Projekts waren derart zeitintensiv, dass es sich nie rein in unserer Schulzeit verwirklichen hätte lassen. Obwohl wir also ca. 1000 Stunden unserer Freizeit in das Projekt gesteckt haben, haben wir es nicht bereut, nicht zuletzt, weil es uns einfach Spaß machte, uns auf neuem, gerade erst von der aktuellen Forschung beachtetem Territorium zu bewegen. Auch die Atmosphäre innerhalb der Gruppe und mit unserem Betreuer und unseren Kooperationspartnern war angenehm und motivierend, sodass „Fun“ zu Recht ein eigener Punkt unserer Projektbeschreibung ist.

13. Gender

Durch eine Unterstützung durch „Forschung macht Schule“ können wir auch die genderspezifischen Aspekte unserer Arbeit besonders herausarbeiten. Obwohl wir der Meinung sind, dass Frauen im Berufsleben nicht aufgrund ihres Frauseins sondern aufgrund ihrer Leistung beurteilt werden sollten (und daher derartige Fragestellungen in einem informationstechnischem Projekt-Bericht gar nicht vorkommen sollten), ist unsere heutige Gesellschaft noch weit davon entfernt: erst kürzlich war Österreich wieder auf vorletzter Stelle der EU bezüglich Einkommensschere zwischen Frau und Mann. Solange diese (sichtbare und unsichtbare) Benachteiligung der Hälfte der Bevölkerung stillschweigend geduldet wird, ist es unserer Meinung nach die Pflicht erfolgreicher Frauen (die ihre Möglichkeit zur Berufsausübung in einer männerdominierten Wirtschaft nicht zuletzt ihren Vorkämpferinnen vergangener Jahrzehnte verdanken) sich ständig für eine Gleichstellung von Frau und Mann in allen Lebensbereichen einzusetzen und nicht als „Alibifrauen“ vom System zu profitieren.

14. Team

Projektleiterin: HAIDER Jasmin, Projektdesign, Programmierung

Projektmanagerin: HEMER Sinja, Projektmanagement, Marketing, Programmierung

Projektbetreuer: RYBIN Erwin, Dr., MBA, etc., Ansprechperson bei Problemen



15. Kooperation

Durch den Erfolg bei Cyberschool.at 2008 und dem dazugehörigen Gewinn des APA Sciene@School Awards wurde einer der Firmengründer von „SIS – Smart Information Systems“, Markus Linder, auf unser Projekt „Advanced Tag Cloud“ aufmerksam. Für eine Weiterentwicklung in Richtung „Semantic Tag Cloud“ werden wir dankenswerterweise von SIS mit Know-how, teilweise von den Unternehmensgründern selbst, unterstützt.

Smart Information Systems entwickelt innovative Lösungen für Produktsuche und Produktberatung im Internet. Das Unternehmen wurde Mitte 2005 von Markus Linder, Martin Schliefnig, Christian Weiss und Svetlana Hollerer in Wien gegründet. (http://www.smart-infosys.com/)

16. Förderungen

Neben der moralischen Unterstützung durch Betreuer, Schule und Elternhaus und der Unterstützung von Know-how durch unseren Kooperationspartner SIS bilden vor allem die Förderungen diverser Stellen einen Rückhalt und vor allem zusätzliche Motivation in schwierigen Momenten des Projekts. Wir danken folgenden Institutionen für Ihre Unterstützung:

Cyberschool.at : Durch die Preise des Jahres 2008 wurden wir ermutigt uns auch heuer in diesem schwierigem Forschungsfeld neuerlich zu betätigen

generation-innovation.at: Durch die großzügige Unterstützung eines Forschungsschecks aus „Forschung macht Schule“ können wir auch die genderspezifischen Aspekte unserer Arbeit besonders herausarbeiten

jugendinnovativ.at: unterstützte dieses Projekt in der schwierigen Anfangsphase durch eine beträchtliche Förderung unserer Projektausgaben