Generierung von ortsbezogenen Informationen zur Darstellung in Schlagwortwolken

Tobias DAHINDEN[a,1] , Daniel EGGERT[a,1] und Oliver FLOHR[b]
[a] *Institut für Kartographie und Geoinformatik, Leibniz Universität Hannover*
[b] *Studiengang Informatik, Leibniz Universität Hannover*

Zusammenfassung. Aus verorteten Texten lassen sich ortsbezogene Informationen ableiten. Eine Möglichkeit diese darzustellen sind Schlagwortwolken. In diesem Artikel wird ein System beschrieben, das solche Informationen im Web abfragt und als Schlagwortwolke darstellt. Der Schwerpunkt liegt auf der Filterung der verorteten Texte. Als verortete Texte werden Wikipediaartikel mit Koordinateneintrag verwendet. Das System wurde auf einem Tablet Computer mit Android Betriebssystem getestet. Das vorrangige Ziel ist es relevante und sinnvolle Schlagworte für eine bestimmte Region zur Verfügung zu stellen. Das wird dadurch erreicht, dass verschiedene Filter-Gewichtungstechniken eingesetzt werden.

Schlüsselwörter. Internet/Web, Extraktion, Wiederfinden, Mobil

Einleitung

Mittels Schlagwortwolken (engl. *tag clouds*) lassen sich Texte stichwortartig und gewichtet darstellen. Diese Art der Informationsdarstellung hat im Web in jüngster Zeit große Verwendung gefunden, zum Beispiel durch Flickr [1]. Für Designzwecke wurde aber schon im Verlauf des 20. Jahrhunderts intensiv Gebrauch davon gemacht [2]. Besonders bekannt ist das Projekt *Wordle* [3], da man dort eigene Texte in ansprechender Form darstellen lassen kann. Ein Beispiel einer Schlagwortwolke, die mit Wordle erstellt wurde, ist in Abbildung 1 zu sehen.

Abbildung 1. Wordle Cloud mit diesem Artikel als Grundlage.

[1] Korrespondenz Adresse: Institut für Kartographie und Geoinformatik, Appelstraße 9a, 30167 Hannover, Deutschland; E-Mail: {tobias.dahinden;daniel.eggert}@ikg.uni-hannover.de.

Wenn man einen Text zu einem bestimmten Ort als Grundlage für die Erstellung der Schlagwortwolke verwendet, spricht man von einer ortsbezogenen Schlagwortwolke. Für solche ortsbezogenen Schlagwortwolken wurden bereits unterschiedliche Textquellen aus dem Internet verwendet: Holenstein und Purves [4] verwendeten Flickr, Paelke et al. [5] benutzten in ihrer Forschung *Wikipedia* und Hahmann und Burkhardt [6] prozessierten den Inhalt von *Openstreetmap*. Als weitere Quellen kämen u. a. *Google Places*, *delicious* und *GeoRSS* in Frage.

Ungeachtet davon welche Quelle verwendet wird, müssen die Texte auf Basis der einzelnen Wörter gefiltert werden, um nicht-relevante Wörter zu eliminieren. Aus den gefilterten Wortlisten können dann Schlagwortwolken generiert werden. Das kann über die API von Wordle erfolgen, wenn das Resultat auf einer Webseite mit Java dargestellt werden soll. Es gibt aber auch Programmierbibliotheken, die die Generierung unterstützen, etwa die Open Cloud Java Library [7].

Bei Schlagwortwolken werden verschiedene Layouts unterschieden [8]. Abbildung 2 zeigt beispielhaft folgende Darstellungen:

a) Sequentielles Layout: ein Wort wird nach dem anderen dargestellt, die Ordnung kann alphabetisch sein oder nach der Häufigkeit der Wörter.
b) Zirkuläres Layout: das ordinal wichtigste Wort steht in der Mitte.
c) Themencluster-basiertes Layout: die Wörter sind nach Themenbereichen gruppiert.
d) Referenzdarstellung.

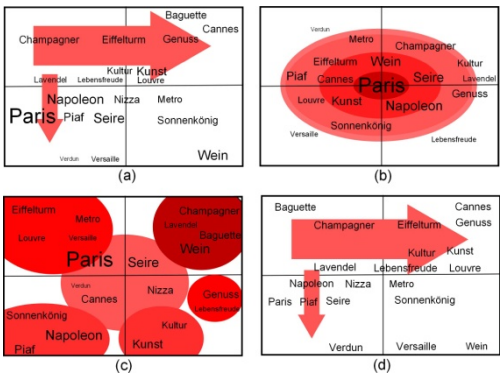

Abbildung 2: Layoutbeispiele für Schlagwortwolken: **a)** sequentiell (alphabetische Sortierung), **b)** zirkulär (abnehmende Popularität), **c)** Cluster (Themencluster), **d)** Referenzdarstellung (sequentiell, ohne Gewichtung) nach [8].

Im Rahmen dieser Arbeit wurden die Filterfunktionen eines bestehenden Systems zur Schlagwortwolkengenerierung [5] erweitert. Als ortsbezogene Texte werden Wikipediaartikel mit Koordinateneinträgen verwendet.

Das Fernziel könnte sein, die so gewonnene Information für die Anreicherung von Routenberschreibungen zu verwenden oder sie als stichwortartige Grundlage für einen neuen Text über ein bestimmtes Gebiet zu benutzen.

1. Filterung von Text

Um eine Schlagwortwolke für einen bestimmten Ort zu erzeugen, werden Texte benötigt, die mit der Umgebung dieses Ortes in einer Beziehung stehen. Eine

Möglichkeit wäre es, über Orts- und Objektnamen im Text einen Ortsbezug herzustellen. Nützlich ist es jedoch, wenn neben dem Text bereits explizit ein Ortsbezug, d. h. Koordinaten, angegeben werden. Der Ort, den die Koordinaten bezeichnen, wird Fußpunkt genannt. Im Prinzip könnte man auch den Ort, an dem ein Text gelagert wird (z. B. eine Bibliothek) als Ortsbezug ansehen. Allerdings kann man dann nicht davon ausgehen, dass der Text diesen Ort beschreibt.

Texte mit Ortsbezug werden folgendermaßen analysiert: Für jeden Text wird in einer Liste festgehalten, wie oft welches Wort enthalten ist. Die Listen der verschiedenen Texte werden gewichtet summiert. Das Gewicht kann beispielsweise invers zur Entfernung des Artikels (Fußpunkt) vom Suchmittelpunkt gewählt werden. Diese Gewichtung ist dann sinnvoll, wenn die Schlagwortwolke für einen Bereich erstellt werden soll, der keine deutliche Begrenzung besitzt.

Abbildung 3 zeigt den Einfluss einer Distanzgewichtung auf die Relevanz der Schlagworte am Beispiel der Schlagwortwolke für die Ortschaft Gleidingen. Während im Bild links zwei Artikel gleichgewichtet sind (einer über Gleidingen und einer über eine Umformeranlage), wurde rechts die inverse Distanz vom Ortszentrum zum Fußpunkt der Artikel gewählt. Der Artikel über die Umformeranlage erhielt dadurch ein kleineres Gewicht. Deshalb fallen rechts Begriffe wie „Umformer“ oder „Kraftwerk“ deutlich weniger ins Gewicht.

Abbildung 3. Einfluss der Distanzgewichtung: Bei der Schlagwortwolke links wurden sämtliche Artikel in der Umgebung gleichgewichtet, rechts wurden hingegen entfernte Artikel weniger gewichtet.

In einem weiteren Schritt wird die Morphologie der Wörter berücksichtigt. Ohne das würden Plural- und Singularformen (z. B. Flüsse, Fluss) als unterschiedliche Wörter aufgefasst. Zur Vereinigung dieser Wörter wird ein Wortstammfilter angewendet.

Weiter müssen aus der Liste die sog. *Stop*-Wörter, das sind besonders häufige Wörter wie „der“, „und“, „ein“, „eine“, etc., und quellenspezifischen Wörter (bei Wikipedia z. B. „bearbeiten“) entfernt werden. Dies erfolgt mit einem *Blacklist*-Filter. Eine Blacklist kann man automatisch erstellen, in dem man die Wörter sämtlicher oder eine große, zufällige Auswahl der zur Verfügung stehenden Texte analysiert und die häufigsten Wörter dieser Liste hinzufügt.

Aus rechentechnischen Gründen findet zuerst die Filterung der Wörter anhand der Blacklist statt, danach wird der Wortstammfilter ausgeführt und erst im letzten Schritt werden die Wörter gewichtet summiert.

2. Implementierung und Beispiel

Um zu testen, ob sinnvolle Schlagwortwolken erzeugt werden können, wurde ein Prototyp erstellt. Dieser basiert auf einem Client-Server-System, schematisch in Abbildung 4 dargestellt. Als Client fungiert ein Android Tablet Computer oder Smartphone, der Server wurde für Linux umgesetzt. Zur Programmierung der Komponenten wurde Java verwendet. Als Beispiel für eine Sammlung verorteter Texte wurde die deutsche Wikipedia verwendet.

Abbildung 4. Schematischer Ablauf der Verarbeitung.

Zur Generierung der ortsbezogenen Schlagwortwolke gibt ein Benutzer ein Gebiet an, das ihn interessiert. Daraus generiert die Client-Anwendung eine passende Abfrage an den Server. Der Server verarbeitet die Anfrage und schickt dem Client eine Schlagwortwolke in Form von einzelnen Worten mit dazugehörenden Gewichten zurück. Um Bandbreite zu sparen, werden nur die sechzig wichtigsten Begriffe übertragen. Der Client muss diese Antwort visualisieren. In unserer Implementierung lässt sich die Schlagwortwolke bisher in alphabetisch-sequentiellem, ordinal-sequentiellem oder zirkulärem Layout darstellen.

Abbildung 5 zeigt die Visualisierung der Schlagwortwolke für einen Bereich um das Hauptgebäude der Leibniz Universität Hannover. Die Größe der Schlagwortwolke hängt von der Displaygröße ab. Bei kleinen Displays wird versucht, das Bild optimal zu füllen. Bei größeren Displays ist die Größe so gewählt, dass alle sechzig Wörter dargestellt werden können. In derselben Karte können mehrere Schlagwortwolken gleichzeitig dargestellt werden, sofern es die Größe des Clientdisplays zulässt.

Abbildung 5. Ortsbezogene Schlagwortwolke mit zirkulärem Layout für eine Umgebung um das Hauptgebäude der Leibniz Universität Hannover. Im Hintergrund liegt ein Ausschnitt aus *Google Maps*.

In einer Datenbank namens Wikipedia-World sind alle Artikelnamen zusammen mit ihren Koordinaten abgelegt. Der Server sucht in dieser Datenbank die Namen der Artikel für ein bestimmtes Gebiet. Anschließend bezieht er die entsprechenden Artikel aus Wikipedia. Die Artikelinhalte werden gefiltert, die einzelnen Wörter gewichtet und das Ergebnis an den Client zurückgeschickt.

Der Vorteil von Wikipedia als Quelle für ortsbezogene Texte liegt darin, dass Artikel zu geographischen Objekten mit einer Koordinate versehen sind (siehe Abbildung 6). Über die Koordinate kann man Artikel finden, die Objekte in einem bestimmten Gebiet beschreiben. Probleme gibt es jedoch, da linien- und flächenhafte Objekte (Flüsse, Länder) nur durch eine einzelne Punktkoordinate verortet sind. Damit werden diese Artikel nur in Ausnahmefällen gefunden. Zudem ist die Verteilung der Artikel im Raum heterogen.

Abbildung 6. Wikipediaartikel mit Koordinaten (gelb hervorgehoben).

Die Implementierung ist im Moment nicht als Webservice umgesetzt. Es wäre jedoch relativ einfach eine Umsetzung auf dem REST-Ansatz [9] zu implementieren. Der Client könnte mit einer GET-Anfrage die Position übermitteln. Der Server würde eine Repräsentation der Ressource (bspw. in XML) zurücksenden.

Der aktuelle Prototyp steht zudem unter http://bit.ly/tagcloudclient zum Download zur Verfügung.

3. Ausblick

Die grundsätzliche Möglichkeit ortsbezogene Schlagwortwolken zu erstellen wurde bereits von verschiedenen Autoren gezeigt. Der hier vorgestellte Prototyp verbessert die Qualität der Schlagwortwolken in dem Sinne, dass Wörter mit demselben Wortstamm zusammengefasst werden und die Gewichtung der Wörter in Abhängigkeit der Distanz zum Fußpunkt des Textes erstellt wird. Dadurch wird besonders in Gebieten mit wenigen Artikeln die Qualität der Schlagwortwolke verbessert.

Es gibt diverse Funktionen, die dem System hinzugefügt werden könnten. Eine Verbesserung wäre etwa, die Darstellung der Schlagwortwolken in Themencluster aufzuteilen, also die Worte themenbezogen zu gruppieren.

Im Moment werden die Schlagwortwolken nur für einzelne Rechtecke bestimmt. Interessant wäre jedoch, die Schlagwortwolken für beliebige Gebiete oder entlang von Strecken zu erzeugen. Damit ließen sich einerseits die Resultate überprüfen, in dem die Begriffe einer Schlagwortwolke mit dem Namensgut und geographischen Begriffen der Region verglichen werden. Andererseits könnte man so zusätzliche, stichwortartige Informationen über ein bestimmtes Gebiet gewinnen.

References

[1] Flickr. http://www.flickr.com/photo/tags Besucht: 19. Jan. 2012.

[2] F.B. Viégas und M. Wattenberg, Tag Clouds and the Case for Vernacular Visualization, *ACM Interactions*, **15** (4), 49-52, 2008.

[3] Wordle. http://www.wordle.net Besucht: 19. Jan. 2012.

[4] L. Holenstein und R. Purves, Exploring places through user-generated content: Using Flickr tags to describe city cores, *Journal of Spatial Information Science*, **1**, Melbourne, 21-48, 2010.

[5] V. Paelke, T. Dahinden, D. Eggert und J. Mondzech, Location Based Context Awareness Through Tag-Cloud Visualizations, *Proceedings of the Joint International Conference on Theory, Data Handling and Modelling in GeoSpatial Information Science*, **28** (2), 290-295, Hongkong, 2010.

[6] S. Hahmann und D. Burghardt, Maple – a Web Map Service for Verbal Visualisation using Tag Clouds Generated from Map Feature Frequencies, *Advances in Cartography and GIScience*, **1**, Springer, Heidelberg Dordrecht London New York, 3-12, 2011.

[7] Open Cloud Library. http://opencloud.sourceforge.net Besucht: 19. Jan. 2012.

[8] S. Lohmann, J. Ziegler und L. Tetzlaff, Ein Blick in die Wolken: Visuelle Exploration von Tag Clouds, *Mensch & Computer 2009: Grenzenlos frei!?*, Oldenbourg Verlag, S. 303–312, 2009.

[9] R.T. Fielding, *Architectural Styles and the Design of Networkbased Software Architectures*. University of California, Irvine, 2000.