histoGraph: Graphbasierte Exploration und Crowdbasierte Indexierung

Wieneke, Lars
lars.wieneke@cvce.eu
CVCE Luxembourg, Luxemburg

Düring, Marten
marten.duering@cvce.eu
CVCE Luxembourg, Luxemburg

Guido, Daniele
daniele.guido@cvce.eu
CVCE Luxembourg, Luxemburg

histoGraph

Der Vortrag wird das im CVCE DH Lab entwickelte Werkzeug histoGraph vorstellen, das die graphbasierte Exploration von digitalisierten Quellen mit crowdbasierter Indexierung verknüpft. histoGraph basiert auf einer zu Demonstrationszwecken entwickelten Software, die Teil des FP7-geförderten Projekts CUbRIK zur Mensch-Maschine-Interaktion in der Multimediasuche war. Der Vortrag enthält neben einer Präsentation des neu entwickelten Designs und des weiterentwickelten Konzepts auch eine Live-Demo. histoGraph wird ab dem Frühjahr 2016 als open source Software frei verfügbar sein.

Mit histoGraph eröffnen wir neue Perspektiven auf die umfangreichen Bestände des Centre Virtuel de la Connaissance sur l'Europe . Gegenwärtig sind dort ca. 20.000 Texte, Bilder und Fotos online verfügbar, hierarchisch organisiert in thematischen Sammlungen (*ePublications*). Diese Sammlungen erzählen die Geschichte der europäischen Integration seit 1945 anhand von sorgfältig ausgewählten Primärquellen.

Exploration

histoGraph ergänzt diese expertenbasierten Sammlungen um einen freieren, explorativen Zugang: Nutzer entscheiden, welche Entität – in unserem Falle: welche Person, Institution oder welches Dokument für sie von Interesse ist.

Das histoGraph-Interface ist in drei vertikale Spalten gegliedert: Die erste Spalte gibt einen ersten Überblick zu seiner Biographie und kookkurrienden anderen Personen. Die zweite Spalte listet alle assoziierten Dokumente auf. Die dritte Spalte repräsentiert diese auf Kookkurrenz basierenden Beziehungen als Graph. histoGraph bietet Nutzern nun mehrere Optionen, diese Ergebnisse zu filtern oder zu sortieren. Von besonderer Bedeutung ist aber die Möglichkeit, gezielt nach Beziehungen zwischen bestimmten Personen zu suchen. Hierzu werden zwei oder mehrere Personen ausgewählt und alle Dokumente aufgelistet, in denen beide erwähnt werden. Darüber hinaus zeigt der Graph alle weiteren Personen, die gemeinsam mit den Gesuchten erwähnt werden. Diese Art der Suche ist inspiriert vom Prinzip des *shortest path*, einer gängigen Methode zur Beschreibung von Netzwerktopologien und -zentralitäten. Hierbei werden alle Schritte gezählt die nötig sind um von einem Knoten des Netzwerks zu einem anderen zu gelangen.

Diese Abfrage funktioniert übrigens ebenso gut für Dokumente oder Institutionen, nur dass in diesem Falle ähnliche Dokumente oder häufig zusammen erwähnte Institutionen dargestellt werden. Dieser Ansatz kombiniert eine gezielte Suche mit einem freieren Finden, dass unerwartete Querverbindungen außerhalb der ursprünglichen Suche sichtbar machen kann. Der größte Unterschied zwischen den eingangs erwähnten hierarchisch organisierten thematischen Sammlungen und histoGraph ist, dass Nutzer die Freiheit haben, ihren eigenen Interessen zu folgen und selbstständig nach für sie relevanten Dokumente und Sozialbeziehungen zu forschen. In histoGraph werden damit mehrere Aspekte historischen Arbeitens aufgegriffen: (1) das genaue Studium einzelner Objekte, (2) deren Betrachtung innerhalb ihres jeweiligen Kontexts, (3) die Suche nach weiteführenden, bislang unberücksichtigten Dokumenten. Die enge Verbindung zwischen Dokumenten und abstrakter Visualisierung sorgt dafür, dass letztere mit Gewinn „gelesen" und evaluiert werden können.

histoGraph arbeitet momentan ausschließlich mit Kookkurrenzen. Es ist mit diesem Ansatz nur sehr schwer möglich, weitergehende Aussagen über die Bedeutung einer solchen Beziehung beispielsweise zwischen zwei gemeinsam erwähnten Personen zu machen: Diese können miteinander interagiert haben, in unterschiedlichen Kontexten erwähnt worden sein oder gar mit dem Hinweis, dass sie absolut nichts miteinander zu tun hatten. Diese Beliebigkeit ist allerdings auch eine Stärke: Sie überlässt Nutzern die Entscheidung, was als eine relevante Beziehung zu gelten hat. Hierbei gilt: Je genauer Beziehungen definiert sind, desto geringer ist der Anteil an irrelevanten Beziehungen. Aber auch: Je großzügiger Beziehungen definiert sind, desto höher ist die Chance, forschungsrelevante Querverbindungen zu entdecken. Im Entwicklungsprozess versuchen wir, die Balance zwischen diesen beiden erstrebenswerten und doch entgegengesetzten Polen zu halten.

Indexierung

histoGraph eignet sich allerdings nicht nur für die Erforschung von digitalen Sammlungen sondern auch für deren Indexierung. Wir arbeiten mit einer Kombination aus

unterschiedlichen Werkzeugen für die Identifizierung von *named entities* wie Personen, Institutionen, Zeitangaben und Orten. Um diese automatisch generierten Annotationen zu prüfen und gegebenenfalls zu verbessern, arbeiten wir zusätzlich mit Methoden des *crowdsourcing*. Hierbei werden einfache Aufgaben, wie etwa die Erkennung von Gesichtern in Fotos oder die Bestätigung eines Datums von so genannten generischen *crowds* übernommen. Anspruchsvollere Aufgaben, wie etwa der Umgang mit Namensvettern bleibt einer *crowd* von Experten vorbehalten. Das System eignet sich ebenso für das kollaborative Indexieren und Annotieren in Teams, etwa einer Projektgruppe.

Im Vergleich mit den bisherigen Sammlungen ermöglicht histoGraph also eine freie Exploration des Materials und das effektive Finden von potentiell relevanten Dokumenten und Beziehungen. Im Zentrum steht hierbei nicht die von Experten kuratierte Auswahl, die mit einem Museumsbesuch vergleichbar ist sondern ein mehr oder minder zielgerichtetes Stöbern, dass einem Archivbesuch näher kommt.

Bibliographie

Centre Virtuel de la Connaissace sur l'Europe (2004-2016), Luxembourg http://www.cvce.eu/ [letzter Zugriff 09. Januar 2016].

Wieneke, Lars / Düring, Marten / Silaume, Ghislain / Lallemand, Carine / Croce, Vincenzo / Lazzarro, Marilena / Nucci, Francesco u. a. (2014): “histoGraph – A Visualization Tool for Collaborative Analysis of Historical Social Networks from Multimedia Collections”, in *Proceedings of 18th International Conference Information Visualisation (IV), 2014 Conference*, Paris.