Interlinking Articles within Online Encyclopedias

Ilire Hasani-Mavriqi
IICM, Graz University of Technology, Austria
ihasani@iicm.edu

Web Science and Web Technology
SS 2011
Guest Lecture
Motivation

• What do you usually do when you need to find a certain information?
Motivation

Just Because Google Exists Doesn’t Mean You Should Stop Asking People Things

Motivation
Motivation

• Accessing web-based information systems such as online encyclopedias is driven by user information needs

• Bounce rate is high [2]

• A bounce occurs when a web site visitor only views a single page on a website, that is, the visitor leaves a site without visiting any other pages before a specified session-timeout occurs [3]
Motivation

- How to tailor the web model of such systems to user needs?
Motivation

• Navigation tools play a certain role
• An article is the basic entry in wiki-based encyclopedias
  – defines and describes a concept or an entity
• Hyperlinks guide readers to pages that provide additional information about concepts mentioned in an article
Austria-Forum

- Three main lexica: AEIOU, Wissensammlungen, and Community
- Lack of links which connect semantically related articles
- Isolated parts such as AEIOU
Kurt Regschek wurde am 29. Juni 1923 als Sohn eines Bankbeamten in Wien geboren, wo er auch aufwuchs.

Nach dem Besuch von Volksschule und Gymnasium legte er Anfang Juni 1940 die sogenannte "Kriegsmatura" ab, wurde zum Jägerbataillon in Berchtesgaden eingezogen und machte den Krieg im Kaukasus und in Afrika mit.

1943 erhielt Kurt Regschek Studienurlaub für die Akademie der schönen Künste in München und wurde hier wegen "Fraternisierens mit einer Französin" des Landesverwaltungsgerichtes angeklagt und saß acht Monate Untersuchungshaft ab.

Ende 1943 wurde er wieder an die Front geschickt, geriet in britische Gefangenschaft und kam dann bis Kriegsende in die Französische Fremdenlegion.


1951 ging er wieder nach Wien zurück, heiratete und bekam eine Tochter (* die Ehe ging bald auseinander, die Tochter starb jung an einem Gehirntumor). 1955 studierte Kurt Regschek als Gast an der Wiener Kunstkademie bei Robin Christian Andersen und Herbert Boeckl.

In all diesen Jahren lebte Regschek von Gelegenheitsarbeiten, als Gebrauchsgraphiker, Skilehrer sowie Graphiker und Statist beim Film, wo er u.a. auch Fritz Muliar kenne lernte.
Regschek’s digitalized work

Index

- Akt
- Bibel
- Biographie
- Freimaurerei
- Frühwerk
- Landschaft
- Liebende
- Mutanten
- Mythologie
- Phantastik
Der Künstler war mit großer Hingabe bemüht, sein reiches handwerkliches und gestalterisches Wissen an seine lehnsbegierigen Schüler weiterzugeben. Die Arbeit während der von ihm veranstalteten Malwochen (Landschaft und Akt) lief sehr diszipliniert ab, wenn auch die Abende lang und gemütlich waren.

Mitte der 80er-Jahre begann für Regschek eine interessante und anregende Lehrtätigkeit an einigen Sommerakademien.


»Es traf sich deshalb gut, dass Regschek in der Eifellandschaft Elemente vorfand, die auch in der Landschaft seiner österreichischen Heimat anzutreffen sind, Berge, Wiesen, bewaldete Hänge, Flusstäler, Schluchten und mächtige Burgen. In der Tat fühlte sich der Künstler schon bald von der Eifel angesprochen, so dass er auch außerhalb einer Lehrtätigkeit hierher zurückkehrte, zumal ihm die Ruhe des ehemaligen Steffeler Pfarrhofes ein ausgezeichnetes Arbeitsklima bot. Inzwischen hat er um die drei Eifelbilder gemalt, die er nun unter der Überschrift »Die Eifel im Spiegel der Jahreszeiten« ausstellte.« (Otto Baur)
Possible Solutions?

- Add links manually?

- Very tedious work – over 81,000 articles
Automatic Linking

• Automatically interlink concepts from an article to articles which describe these concepts
• Provide high quality links within the online encyclopedia
• Overcome the problem of isolated lexica
High Quality Links

1. High Quality Links
2. Enhance Network Navigability
3. Attract users to follow links
4. Acquisition of Networked Knowledge
Networks (1/2)

• Graphs are mathematical structures consisting of vertices and edges connecting the vertices

• When we observe large graphs that exist in nature, societies, or systems we refer to them as networks

• Networks consist of nodes and links which connect those nodes (computer science)
Networks (2/2)

- **Social networks**: nodes are people and links are acquaintances, friendship, and so on.
- **Communication networks**: Internet: nodes are computers and links are cables connecting computers.
- **Biological networks**: Metabolism: nodes are substances and links are metabolic reactions.
- **Information networks**: Web: nodes are Web pages and links are hyperlinks connecting pages.
Research in network navigability initiated by Miligram’s small word experiment
Participants were only allowed to send the letter to their friends
Letters that reached the destination needed on average about 6 hops
Population of the USA constituted a “small world”.

Network Navigability (1/5)
Network Navigability (2/5)

From the structural point of view:

• *Network is navigable if and only if there is a short path between all or almost all pairs of nodes in the network.*

Formally:

• There exist a giant component – containing almost all nodes

• The effective diameter is low-bounded by \( \log(n) \), where \( n \) is the number of nodes in the network
Network Navigability (3/5)

Example 1

Network is not navigable because there is no giant component, i.e. the network is not connected
Network Navigability (4/5)

Example 2

Now, there is a giant component, i.e. the network is connected. However the network is not navigable because effective diameter = 6, and $6 > \log_2(8)$
Example 3

The network is navigable because there is a giant component and effective diameter = 2 which is bounded by $\log_2(10)$
Local Network Navigability (1/4)

• Previous examples – global network navigability- centralized search

• A network is navigable but we possess only local knowledge of network

• When we arrive at a particular node we know only outgoing links from that node and nothing beyond that

• Procedures to find an arbitrary target node from an arbitrary start node - decentralized search
Local Network Navigability (2/4)

- Search in social networks is a typical example of decentralized search.
- Structural network properties which enable decentralized search.
- Local network navigability requires:
  - Existence of network hubs that are connected to many nodes.
  - Existence of network clusters where nodes are highly interlinked.
Local Network Navigability (3/4)

1. Power-low degree distribution with exponent $\gamma$
2. High clustering coefficient $C$
3. External to the network: nodes similarity

A is start node and D is target node
B is selected based on degree.
There are seven possible paths from B. The optimal path leads to C - clustering
Additional hint that can guide us in selecting C over E is the node similarity
Local Network Navigability (4/4)

• Degrees, clustering, similarity - background knowledge about the network
• Network background knowledge guides us in our search for a target node
• More than one link to follow - consult the background knowledge
  – which of the links will lead us with highest probability to a given target node?
Fragmentation of Knowledge

• Readers skip irrelevant parts of an article
• Only small fragments of an article are read
• Learning process ends with acquisition of fragmented knowledge
• Context, correlations, or connections between parts of knowledge are not visible
Networked Knowledge

- Hypertext as knowledge base
- Node interlinkage
- Fully connected network of articles
- Users understand in depth concepts invoked
- Networked knowledge is gained
Automatically Generated Links

• **Link detection**
  determining which terms or phrases of an article could serve as anchors

• **Link disambiguation**
  specifying which articles to link to
Link Detection

• Also known as keyword extraction [4]
• Identify important words and phrases within an article
  – technical terms
  – named entities
  – new terminology
  – other concepts closely related to the content of the article
Link Disambiguation

• Also known as word sense disambiguation [4]
• Find the correct article that should be linked to a candidate keyword
• Problem of link ambiguity
• One word can be linked to more than one article
• Find the correct sense of the word
  – context where it occurs
Link Quality

• **Misslinking** – a term or phrase is linked to an incorrect link target

• **Overlinking** – a term or phrase is linked when there should be no link at all

• **Underlinking** – a term or phrase is not linked when there should be a link [5]
Wikiﬁcation

• Automatic linking of web resources
  – within a knowledge base
  – to Wikipedia
  – on the fly
Linking Concepts within a Knowledge Base

• NNexus [5]

• automatically links entries of a collaborative online encyclopedia PlanetMath² into a semantic network of concepts using metadata of the entries

• Special content – strict classification

• Classification mechanism and a set of linking policies applied first (++) precision and recall

² http://planetmath.org/
Linking Concepts to Wikipedia

• Wikify! [4]
• Learning to link with Wikipedia [6]
• Automatic link detection: a sequence labeling approach [7]
• Manually linked pages in Wikipedia are highly accurate
• Good training set for machine learning based automatic linking
Linking Concepts on the fly

- Linkator [8]
- Information extraction mechanisms +
- Semantic Web technologies
- Appropriate target link determined with LinkedData³
- Link destination is semantically determined, on the fly

³http://linkeddata.org/
Automatic Linking in Austria Forum

- Node interlinkage within the knowledge base
- Evaluate the navigational effectiveness of automatic links from the network-theoretic perspective
- Semantic evaluation of the quality of linking
Network Navigability Evaluation Framework

• Decentralized search as a mechanism to evaluate navigational effectiveness

• First phase:
  – Measure different structural network properties
  – Conclude that a particular network is navigable or on-navigable
Calculating Network Navigability

• By placing a network into $\gamma$, C space we get only the answer to the question: is the network locally navigable?
• We can not quantify the degree of navigability
• Second phase:
  – Simulate navigation in a network
  – We can empirically refine our measurement results
  – We can measure to which extent a network is navigable
Navigability Evaluation of Automatic Links

• Early version of Austria-Forum ~ 45,000 articles

• Network Generation
  – For each article we select the top three concepts to create links.
  – The final network has 48,170 nodes and 173,608 links
Navigability Evaluation of Automatic Links

- Connectedness and Diameter

- The network is connected (a) – the largest strongly connected component has 85.19% of nodes (lower navigability bound)
- The effective diameter (b) is 5.97 and is bounded by log(n)

Therefore, we can conclude that this network is navigable and proceed to the next step.
Navigability Evaluation of Automatic Links

- Degree Heterogeneity and Clustering

- Calculate the distributions of in-degree (a), out-degree (b), and clustering-coefficient and estimate power-law exponents $\gamma$ of the degree distributions ($\gamma = 1.82$ for in-degree $\gamma = 1.79$ for out-degree)

- The lower bound for clustering-coefficient in this case is about $C > 0.2$.

- We obtained $C = 0.26$, which is very close to the lower bound
Navigability Evaluation of Automatic Links

• Greedy Navigation
• We simulate greedy navigation in the network by choosing randomly 1,000,000 node pairs
• Each simulation begins at a start node and heads towards a destination node
• Measure the success rate of the greedy navigation and its stretch i.e. the ratio of the number of greedy hops to the global shortest path
• Success rate very low – 3.7%
Combining Automatic Links and Breadcrumbs

• Network Generation - 48,229 nodes and 333,624 links.

• Connectedness and Diameter
  – The network is completely connected and the largest strongly connected component has 99.99% of nodes. The effective diameter is 2.9. The network is navigable.

• Degree Heterogeneity and Clustering
  – $\gamma = 1.86$ for in-degree and $\gamma = 1.77$ for out-degree distribution. Clustering-coefficient $C = 0.63$ – network is efficiently navigable

• Greedy Navigation – success rate 100%
Semantic Evaluation of the Quality of Linking

• Semiautomatic Linking Tool
Linking Tool

Diagram details:
- End User
  - wiki article
  - filter/remove markup
  - extract and rank candidates
  - calculate context
  - integrate suggested links
  - find best links with Lucene
  - construct Lucene queries
- wiki article with integrated links
Filtering

- Filtering
  - remove wiki markup, stop words
  - aggregate n-grams
  - pass them to the next stage
Extract Best Candidates

- Term Frequency/Inverse Document Frequency or tf.idf

- *Number of occurrences of a term in a given document multiplied with the (often logsmoothed) inverse of the number of documents where the term appears* [4]

\[
\text{tf.idf}_{t,d} = (0.5 + 0.5* \frac{f_{t,d}}{\max_r(f_{r,d})}) \times \log(N/n_t)
\]

where \(f_{t,d}\) is the frequency of term t in a document d

\(\max_r(f_{r,d})\) is the maximal frequency of any term in d

N is the total number of documents and \(n_t\) is the number of documents that contain the term t [9] [10]
Find Link Suggestions

• Query engine of Apache Lucene⁴ is employed to determine link destinations

• Calculate context
  – context in which a particular word appears is very important to discover the meaning of the word
  – the context i.e. the sentence where the word appears is stored and used when constructing Lucene queries

• Queries based on: article’s title, content, and word’s context

• Integrate best link suggestions

⁴ http://lucene.apache.org/
Semantic Evaluation

• Linking Tool and three members of editorial team
• “Mimic" content authors in order to enhance the quality of automatically generated links
• Data set:
  – 109 randomly selected articles from Austria-Forum (30% very short)
  – 545 identified concepts which could serve as link anchors
Semantic Evaluation

• Concepts are automatically detected and highlighted by the Linking Tool - link disambiguation is evaluated

• From all 545 selected key concepts only for 9 of them, editors were convinced that there exists another link destination

• Precision is defined as the number of correctly created links divided by the number of all invoked links by editors – 96% precision
Future Work

• Links created by editors - golden standard for future work, i.e. machine learning approaches could be applied
• Complement the algorithm with a set of policies derived from evaluation data
• Evaluate link detection
• Replace simulations with actual user click-data
• Investigate the effects of different navigational tools on the Wikipedia dataset
References (1/2)


Thank you for your attention!

Questions?