Information Retrieval in Blogs

Talk in winter semester 07/08
Advanced seminar „Information Retrieval“ (PD Dr. Karin Haenelt)

21.01.2008

Sabrina Lin, Shajy Valiath, Torsten Hopp
Outline

1. Basics
   • What is a blog?
   • Which features provides a blog?
   • Examples

2. Blog search engines
   • Why do we need a blog search engine?
   • What blog search engines are there?

3. Blog Task at the TREC 2006
   • Challenge and composition of the track
   • Approaches
   • Results

4. Summary & Outlook
1. Basics ► What’s blog all about?

- Short form for web log
- Journal style websites
- To share ideas among software communities
- Enables to publish wide variety of contents to web

References: [Lan07]
Jorn Barger has called these journal like websites as Weblogs

References: [Tec06], [Wik08a], [Blo00]
1. Basics ► History

- The word weblog has again shortened to Blogs by Peter Merholz

References: [Tec06], [Cor07], [Blo00]
1. Basics ► History

- AP wire similar to large chat room electronic conversation
- „Amateur radio“ or „Ham radio“
- Usenet, email lists, Bulletin Board Systems (BBS)
- Internet forum software such as WebEx
- With the evolution of www, everyone wanted his/her presence in Internet

References: [Tec06], [Blo00]
1. Basics ► Features

- A way of sharing knowledge
- No need to be a programmer
- WYS / WYG
- Organize over time
- Connecting other people
- Web log go beyond email
- Easy to maintain
- Search engine friendly

References: [Lan07], [Wor07]
1. Basics ► Technologies

- Client-server architecture
- In the server runs a webserver
- PHP and MySQL
- The blogsoftware is in PHP implemented
- The blog entries are saved in Database
- The Client (here the user) uses a Web browser to access the data

References: [Lan07], [Wor07]
### RSS Feeds

- RSS stands for „**Really Simple Syndication**“

- The XML-based file in which the blog hosting software places a machine-readable version of the blog so that it may be „Syndicated“ for further distribution on the web. Formats such as RSS and Atom are used to structure the XML file.

References: [Sof07]
1. Basics ▶ RSS Feeds

- It’s way to easily distribute a list of headlines, update notices and sometimes content to a wide number of people.

- It’s used by computer programs that organize these headlines notices for easy reading.

References: [Sof07]
What information does RSS provide?
- For example, the RSS information for headlines on a local news website could contain the following information:

<table>
<thead>
<tr>
<th>Item 1:</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Title:</strong> Sidewalk contract awarded</td>
</tr>
<tr>
<td><strong>Description:</strong> The city awarded the sidewalk contract to Smith Associates. This hotly contested deal is worth $1.2 million.</td>
</tr>
<tr>
<td><strong>Link:</strong> <a href="http://www.gardencitynews.com/contractawards/sidewalk.htm">http://www.gardencitynews.com/contractawards/sidewalk.htm</a></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Item 2:</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Title:</strong> Governor to visit</td>
</tr>
<tr>
<td><strong>Description:</strong> The governor is scheduled to visit the city on July 1st. This is the first visit since the election two years ago. The mayor is planning a big reception.</td>
</tr>
<tr>
<td><strong>Link:</strong> <a href="http://www.gardencitynews.com/news/2004/06/gov-visit.htm">http://www.gardencitynews.com/news/2004/06/gov-visit.htm</a></td>
</tr>
</tbody>
</table>

References: [Sof07]
What information does RSS provide?

- RSS provides very basic information to do its notification. It is made up of a list of items presented in order from newest to oldest. Each item usually consists of a simple title describing the item along with a more complete description and a link to a web page with the actual information being described. Sometimes this description is the full information you want to read (such as the content of a weblog post) and sometimes it is just a summary.

References: [Sof07]
1. Basics ► Permalink

- **Permanent Link**
- On a blog site, the front page constantly updates. Permalinks link to that one article forever
- a type of URL designed to refer to a specific information item
- This is important, as most blogs change regularly, and without a permanent link, the posts would be impossible to find later.
- Permalinks are usually indicated by a pound-sign (#) or the permalink at the end of the post.

References: [Lan07], [Wor07]
Kevin Smith's report from the Cannes Film Festival, where Clerks II got an 8-minute standing ovation. Harvey Weinstein: "In my thirty years of coming here, I've never seen a standing ovation last that long at a midnight show in Cannes. Ever." #

"Web 2.OH, YEAAHH!!" tshirts...drink the Kool-Aid!

Feature request: per-domain JavaScript disabling. God yes, any more than one NY Times story up in Safari throws beach balls like crazy. #

US TV schedule for World Cup 2006. Goaaaaaaaaaaalllllllllllll! #

http://www.kottke.org/remainder/06/05/11124.html
1. Basics ► Blogroll

- A list of blogs.
- A blogger features a list of their favorite blogs in the sidebar of their blog.
- http://blog.beetlebum.de

References: [Lan07], [Wor07]
Outline

1. Basics
   - What is a blog?
   - Which features provides a blog?
   - Examples

2. Blog search engines
   - Why do we need a blog search engine?
   - What blog search engines are there?

3. Blog Task at the TREC 2006
   - Challenge and composition of the track
   - Approaches
   - Results

4. Summary & Outlook
2. Blog search engines ► Motivation

- 1,4 Mio. new blog entries everyday

- Index update of „common“ search engines often lasts some days
  - To slow for blogs!

References: [Roe06], [Sif07], [ORMMS06]
2. Blog search engines ► Providers

- Blog search engines
  - General blog search engines
    • E.g. BlogDigger, Technorati, BlogPulse
  - Blog search engines for special subjects
    • E.g. Blawgsearch, Photoblogs.org
- Specialised services of common search engines
  • E.g. Google, Yahoo, AskJeeves
- Blog Directories
  • E.g. Blogorama, BlogUniverse

References: [MBI06], [MO06]
2. Blog search engines ► Technorati

- Founded by David Sifry in 2002
- Principle: „Blogger helps blog search engine“
  - Ping to Technorati when a new article is published
  - Indexing of the article by Technorati
- Article is findable mostly within few minutes
- Ranking-Algorithm:
  - Similar to Google: number of links pointing on the blog entry
  - Count of a link older than six months is discarded → considers only the current status of the blog

References: [Bas07]
2. Blog search engines ► Technorati

- 2006: Redesign
  - Portal-like layout
  - Use of tags

- Search options:
  - Keyword search
  - Advanced search
    - Restriction to bookmarks, blogs with a certain topic, URLs
    - Tag search
    - Search in blog categories

References: [Roe06]
2. Blog search engines ▶ Technorati

Keep track of all your favorite blogs

How? With Favorites! If the blogosphere seems overwhelming, we can help. Tell us what your favorite blogs are, and we'll keep track of them.

Create your Favorites now!

Popular

The most popular news articles bloggers are linking to right now:

- Bobby Fischer, Chess Master, Dies at 64 - nytimes.com
- Chess legend Fischer dies at 64 - news.bbc.co.uk
- Confederate Flag Takes Center Stage Once Again - nytimes.com
- Bush Calls for $115 Billion Economic Aid Package - nytimes.com
- Golfweek Fires Editor Over Noose Cover - nytimes.com
- White House Study Found 473 Days of E-Mail Gone - washingtonpost.com
- Mature Human Embryos Created From Adult Skin Cells - washingtonpost.com
- Nasa investigates virtual space - news.bbc.co.uk

Top Tags

"Frank rich", "ron paul", "authority bach bercampcologne2 britney facebook fashion"
2. Blog search engines ► Comparison of results

- Scientific comparison is challenging
  - Is the blog relevant?
  - Is it really a blog? Or a newspaper website?
→ Hard to avoid subjectivity!

Comparison of results from *Technorati* and *Google Blog-Search* for a certain query
- Keyword search for „Lokführer Tarifverhandlungen“ on 19.12.2007
## 2. Blog search engines ➤ Comparison of results

<table>
<thead>
<tr>
<th>Rank</th>
<th>URL</th>
<th>Date</th>
<th>Relev.</th>
<th>Comm.</th>
<th>URL</th>
<th>Date</th>
<th>Relev.</th>
<th>Comm.</th>
</tr>
</thead>
</table>
Outline

1. Basics
   • What is a blog?
   • Which features provides a blog?
   • Examples

2. Blog search engines
   • Why do we need a blog search engine?
   • What blog search engines are there?

3. Blog Task at the TREC 2006
   • Challenge and composition of the track
   • Approaches
   • Results

4. Summary & Outlook
3. TREC 2006 Blog Task: Challenge

- Characteristics of blogs:
  - Mostly the articles are not editorial
  - Splogs (Spam Blogs)
  - Articles express opinions, subjectivity
  - Comments

- Characteristics of queries (intention of users)
  - Often looking for opinions about a certain topic

- TREC 2006 Blog Task: not only keyword search but also search for opinions to a topic given by the keyword

References: [MO06], [ORMMS06]
Corpus should reflect the blogosphere

About 100,000 blogs were observed
  - December 2005 – February 2006
  - Capture of
    - RSS-Feeds
    - HTML-Permalinks (with comments)
    - Complete Website
  - Mainly in English

Quality of blogs is varying

References: [MO06]
3. TREC 2006 Blog Task ► Corpus Composition

- Choice of blogs
  - About 70,000 top-blogs
    • Chosen by Nielsen BuzzMetrics and the University of Amsterdam
  - About 18,000 Splogs
    • Copied content, linkfarms
  - About 12,000 other blogs
    • News, sports, politics, health

- Reorganisation of the collection to track format
- 753,000 Feeds, 3.2 Mio. Permalinks
  - About 140 GB data

References: [MO06], [ORMMS06]
3. TREC 2006 Blog Task ► Corpus Composition

- Identification of relevant documents by jurors
  - Jurors create tasks on the basis of search queries
  - Search in blog track for relevant documents by the TREC-participants → pooling
  - Evaluation of submitted documents in respect to the task by a metric: -1, 0, 1, 2, 3, 4

<top>
<num> Number: 871
<title> cindy sheehan
<desc> Description:
What has been the reaction to Cindy Sheehan and the demonstrations she has been involved in?
<narr> Narrative:
Any favorable or unfavorable opinions of Cindy Sheehan are relevant. Reactions to the anti-war demonstrations she has organized or participated in are also relevant.
</top>

References: [MO06], [Hae03]
3. TREC 2006 Blog Task  ►  Processing

- Manual and automatic queries
- Result: Top 1000 documents expressing an opinion
- 14 participants
  - 57 submitted results
- Measurement for evaluation:
  - MAP (mean average precision)
  - R-precision
  - Binary preference
  - Precision at 10 documents (P@10)

References: [MO06]
Presentation of approaches from two participants:

- BlogVox Opinion Retrieval  
  (University of Maryland)

- Web Information Discovery Integrated Tool (WIDIT)  
  (Indiana University)
3. TREC 2006 Blog Task ► BlogVox Opinion Retrieval

- **Task**

  - To build a system that will take a query string describing a topic, e.g., “March of the Penguins”
  
  - It then returns a ranked list of blog posts that express an opinion, positive or negative, about the topic
  
  - The participants train this system to work on the dataset provided from NIST

References: [JKF06]
Approach:

- First, identify posts from splogs (spam blogs) using a machine-learning based approach.
- Then, eliminate them from the collection.
- Next, index “cleared” posts to eliminate irrelevant text associated with navigation links, blog-rolls, link-rolls, advertisements and sidebars.
- Finally, the system applies a set of scoring modules to rank the result relevancy.

References: [JKF06]
Design-Goal:

“This system should be able to dynamically learn topic sensitive sentiment words to better find blog posts expressing an opinion about a specified topic.”

References: [JKF06]
3. TREC 2006 Blog Task ► BlogVox Opinion Retrieval

- BlogVox Preparation steps:

1. Parse the TREC corpus
2. Remove non English posts
3. Eliminate splogs from the collection
4. Remove false material from the DOM tree

References: [JKF06]
System Overview

References: [JKF06]
Identifying and removing spam

- 2 kinds of spam are common in the blogosphere:
  - Spam blogs or splog
    - Hosting ads
    - Promoting affiliate sites
    - Getting new pages indexed
  - Spam comments
    - A spam inside a blog comment

Shared common purpose

- Indexing new pages
- Promoting their page rank
- Selling online merchandise or hosting advertisement

References: [JKF06]
A typical splog

References: [JKF06]
Detecting splogs

- BlogVox used only local features to detect splogs
- A statistical model is used to detect splogs
  - Based on supervised machine learning techniques
    - Based on a training set
  - Using content local to a page

References: [JKF06]
Identifying post content

- Most of the extraneous features in blog post are links
  - Content-links:
    - Either title or the text of the post
  - Extra-links:
    - Not directly related to the post, but provide additional information, such as *navigational links, blog rolls*...

References: [JKF06]
3. TREC 2006 Blog Task ▶ BlogVox Opinion Retrieval

- A typical blog post

References: [JKF06]
Identifying post content

- Extraneous links tend to be tightly grouped containing relatively small amounts of text.
  - First create a DOM tree (HTML Document Object Model)
  - Then traverse it attempting to eliminate any extraneous links
    - Based on the following procedures:
      A link $a$ is eliminated if another link $b$ within a $\theta$(dist) tag distance exists such that:
      - No title tags (H1, H2..) exist in $\theta$(dist) tag window of $a$
      - Average length of the text bearing nodes between $a$ and $b$ is less then some threshold
      - $b$ is the nearest link node to $a$

$\theta$(dist) of 10 tags was chosen

References: [JKF06]
Scoring

- Ranking the likelihood of expressing an opinion about a given topic
  - Query Word Proximity Scorer
  - Parametrized Proximity Scorer
  - Positive and Negative Scorers
  - Lucene Relevance Score

- First, use SVMs to combine scorers from above functions
  - SVMs were trained using a set of 670 samples of which 238 were positive and the rest were negative.
- Then combine with Lucene relevance score to produce final runs

References: [JKF06]
Results

- Cleaning significantly improved both opinion and retrieval scores of the system.

References: [JKF06]
Conclusion

- The overall performance as measured by MAP and R-precision scores showed that the system worked well on most of the test queries.

- It is believed that the system can be improved by increasing the accuracy of the post-cleaning and refining the opinion scorers.

References: [JKF06]
3. TREC 2006 Blog Task ► WIDIT

- WIDIT - Web Information Discovery Integrated Tool

- The goal of opinion task:
  
  “uncover the public sentiment towards a given entity/target”

- Retrieving topically relevant blogs
- Identifying those that contain opinions about the target
- The blog test collection contains considerable amount of noise

References: [YYVZ06]
Research Focus

- Does noise reduction (e.g., exclusion of non-English blogs, navigational text) improve blog retrieval performance?
- How can on-topic retrieval system be optimized to address the challenges of short queries typical in blog retrieval?
- What are the evidences of subjectiveness / opinion and how can they be leveraged to retrieve opinionated blogs?

References: [YYVZ06]
3. TREC 2006 Blog Task ▶ WIDIT

- **Approach**
  - (Noise reduction)
  - Initial retrieval
    - Standard WIDIT retrieval method
  - On-topic retrieval optimization
    - Post-retrieval reranking approach
  - Opinion identification
    - A fusion of 4 opinion modules

References: [YYVZ06]
WIDIT blog opinion retrieval system architecture

References: [YYVZ06]
Noise reduction

- In order to assess the effect of the noise reduction, two separate indexes were created (with and without noise)

- 2 modules:
  - NBI (*Non-English Blog Identification*)
    - Ex: ä, ö, ü, β, é, ç, Ω, â, œ, μ....
    - Ex: Identification ⇔ Identifikation...
  - BNE (*Blog Noise Elimination*)
    - Use markup tags to differentiate blog content
    - Ex: non-blog content: (sidebar, advertisement, forms, header, footer...)

Blog content: (post, comments, ...)

References: [YYVZ06]
Initial retrieval

- WIDIT retrieval engine
  - Document/query indexing
    - Retrieval module
- Removing stopwords:
  - Non-meaningful word from a standard stopword list
    - a, the, though, of..
  - Non-alphabetical words
  - Word > 25 or word < 3 characters
  - Contain > 3 repeated characters

References: [YYVZ06]
On-topic retrieval optimization

- Reranking the results of the initial retrieval
  - Based on a set of topic-related reranking factors
    - Exact match
    - Proximity match
    - Noun phrase match
    - Non-Rel match

- Purpose:

“to float low ranking relevant document to the top ranks based on post-retrieval analysis of reranking factors”

References: [YYVZ06]
Opinion identification

- Accomplished by combining the 4 opinion modules that leverage various evidences of opinion
- 3 sources of evidence:
  - **Opinion lexicon**: a set of terms often used in expressing opinions
    - Ex: “Skype sucks”, “Skype is cool’”
  - **Opinion collocations**: collocations used to mark adjacent statements as opinion
    - Ex: “I believe God exists”
  - **Opinion morphology**: people often use morphed work form for emphasis, when expressing strong opinions or perspectives
    - Ex: “Skype is soooo buggy”

References: [YYVZ06]
Opinion identification

- 4 modules:
  - Opinion Term Module
    - Using Opinion Term Lexicon, which extracted from training data
  - Rare Term Module
    - Ex: sooo good
  - IU Module
    - “I believe” and “good for you”
  - Adjective-Verb Module
    - “It assumes that a post with a high concentration of subjective adjective and verbs must be opinionated.”
    - PSE set: “Potentially Subjective Element” set

References: [YYVZ06]
Opinion identification

- Modules scores:
  - Opinion Term Module
    - 2 OMT scores are generated: document-length normalized frequency of OT terms in document and OT terms near query string in document
  - Rare Term Module
    - 2 RT scores similar to OT scores are computed.
  - IU Module
    - 2 IUM scores similar to OTM and RTM

References: [YYVZ06]
Fusion

- The fusion module combines the multiple set of search results after retrieval time.
  - Two common fusion formulas
    - Similarity Merge
    - Weighted Sum

- WIDIT employs “the weighted sum formula”
  - Sums the normalized system scores multiplied by system contribution weights.

References: [YYVZ06]
3. TREC 2006 Blog Task ▶ WIDIT

- Results

Table 1. Official TREC blog opinion results of top 5 systems

<table>
<thead>
<tr>
<th>Group</th>
<th>MAP</th>
<th>MRP</th>
<th>P@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indiana University</td>
<td>0.2052</td>
<td>0.2881</td>
<td>0.468</td>
</tr>
<tr>
<td>Univ. of Maryland</td>
<td>0.1887</td>
<td>0.2421</td>
<td>0.378</td>
</tr>
<tr>
<td>Univ. of Illinois at Chicago</td>
<td>0.1885</td>
<td>0.2771</td>
<td>0.512</td>
</tr>
<tr>
<td>Univ. of Amsterdam</td>
<td>0.1795</td>
<td>0.2771</td>
<td>0.464</td>
</tr>
<tr>
<td>Univ. of California, Santa Cruz</td>
<td>0.1549</td>
<td>0.2355</td>
<td>0.438</td>
</tr>
</tbody>
</table>

- **The best** of blog opinion results
- **Assumption:**
  - Noise reduction will improve retrieval performance
- **Result:**
  - Slightly increased **P@10**, but generally decreasing **MAP** and **MRP**
  - **The noise reduction also eliminated some true blog content!**

References: [YYVZ06]
Results: query length effect

- The longer the queries, the better the retrieval result
- Long query may contain description of opinions, that helps finding opinion blogs while retrieving non-topical blogs at the same time.

References: [YYVZ06]
Results: opinion reranking effect

- The opinion reranking improved overall performance
- It clearly demonstrates the effectiveness of WIDIT’s opinion reranking approach

References: [YYVZ06]
Conclusion

- WIDIT's fusion approach of combining multiple sources of evidence and multiple methods worked well for TREC's blog opinion retrieval task.

- Even though each method contributed to improving retrieval performance, combing of 3 methods achieved the best overall performance.

References: [YYVZ06]
Outline

1. Basics
   • What is a blog?
   • Which features provides a blog?
   • Examples

2. Blog search engines
   • Why do we need a blog search engine?
   • What blog search engines are there?

3. Blog Task at the TREC 2006
   • Challenge and composition of the track
   • Approaches
   • Results

4. Summary & Outlook
4. Summary & Outlook

- Increasing number of blogs and blog posts
  - Blog search engines needed to get an up-to-date overview of the blogosphere

- Blogs have special characteristics that have to be considered in Information Retrieval
  - not editorial, spam, opinions, subjectivity, comments

- TREC 2006 Blog Task: Challenging teams for opinion retrieval
  - BlogVox Opinion Retrieval, WIDIT
  - Challenge: noise reduction!

- Special „discipline“ Blogs just started

- Further development in the future!
  - TREC 2008 Blog Track: Call for participation
Thanks for your attention

Questions?


References (3)


ANHANG: Mean average precision

- am häufigsten verwendetes Zusammenfassungsmaß
- Sowohl Recall als auch Precisionorientierte Aspekte
- Berechnung:
  - MAP ist Mittelwert der durchschnittlichen Precision.

1. durchschnittliche Precision für jede Anfrage berechnen (= Average Precision AveP)
2. Mittelwert von AP über alle Anfragen

\[
\text{AveP} = \frac{\sum_{r=1}^{N} (P(r) \times \text{rel}(r))}{\#\text{rel}}
\]

\( r = \) Rang eines Dokuments
\( N = \) Anzahl gefundener Dokumente
\( \#\text{rel} = \) Anzahl relevanter Dokumente
\( \text{rel}(r) = \) binäre Realtion, die Relevanz eines Dokuments angibt
\( P(r) = \) Precision für die Anfrage \( r \)

Quellen: [SL05], [Wik07a]
ANHANG: R-Precision / Precision@10 / Binary Pref.

- **R-Precision**
  - Precision an der R-ten Position im Ranking
  - R: Anzahl relevanter Dokumente

- **Precision@10**
  - Precision bei 10 Dokumenten

- **Binary Preference**
  - Chris Buckley, Ellen Voorhees
  - Summe über bekannte relevante Dokument für ein Thema t und die IR-Methode M.
  - 0 ≤ Bpref(t,M) ≤ 1

\[
B_{pref}(t, M) = \frac{1}{r} \sum_{i=1}^{r} 1 - \frac{N_r(d_i, t, M)}{r'}
\]

*N_r(d_i, t, M): Anzahl bekannter nicht-relevanter Dokumente (von den ersten r nicht relevanten Dokumenten) über Rang d*

Quellen: [SL05], [Grö05]
ANHANG: Post Content Identifying Algorithm

Algorithm 1 Blog post cleaning heuristic

Nodes[] tags = tags in the order of the depth first traversal of the DOM tree
for all i such that 0 ≤ i ≤ |tags| do
    dist = nearestLinkTag(tags, i);
    if dist ≤ θ_dist then
        eliminate tags[i]
    end if
end for

Procedure 2 int nearestLinkTag(Nodes[] tags, int pos)

minDist = |tags|
textNodes = 0
textLength = 0
title = false;
for all j such that pos − θ_dist ≤ j ≤ pos + θ_dist do
    node = tags[j]
    if j = 0 || j = pos || j > (|tags| − 1) then
        continue
    end if
    if node instanceOf TextNode then
        textNodes++;
        textLength += node.getTextLength();
    end if
    dist = |pos − j|
    if node instanceOf ListNode && dist < minDist then
        minDist = dist
    end if
    if node instanceOf TitleNode then
        title = true
    end if
end for
ratio = textLength / textCount
if ratio > α_avgText || title == true then
    return tags.size()
end if
return minDist
Query Word Proximity Scorer:
- “The average number of sentiment terms occurring in the vicinity of the query terms using a window size of 15 words before and after the query term”

Parametrized Proximity Scorer:
- “highly polar sentiment words” and “the relatively less polar words”
- “used parameters to specify the window of text to search for sentiment words (5 and 15), and boost sentiment terms around phrase queries (1 and 3)."
ANHANG: Score Module

- Positive and Negative Scorers:
  - “counted the number of sentiment words (positive and negative) in the entire post”

- Lucene Relevance:
  - “was used to find how closely the post matches the query terms”