Navigating the Intranet with High Precision

Huaiyu Zhu
Alexander Löser
Sriram Raghavan
Shivakumar Vaithyanathan

IBM Almaden Research Center
What is this paper about?

1. Total number of hits in IBM intranet = 4839
2. The acronym IDP has 11 different expansions
3. Top required hit refers to Individual Development Plan

Where to begin: Explore Your Career site has the guidance and format to assist in an effective development discussion and building your development plan.

Valued skills: Access the Market Valued Skills quickview to learn about the skills you need.

Find out more: Descriptions of new IDP features and Frequently Asked Questions.

Terms of use
## Top unique 10 queries on IBM Intranet

<table>
<thead>
<tr>
<th>Number of queries</th>
<th>Query string</th>
<th>Intended expansion of acronyms</th>
<th>Number of Variants</th>
</tr>
</thead>
<tbody>
<tr>
<td>160824</td>
<td>idp</td>
<td>Individual Development Plan</td>
<td>75</td>
</tr>
<tr>
<td>91442</td>
<td>issi</td>
<td>IBM Standard Software Installer</td>
<td>33</td>
</tr>
<tr>
<td>83790</td>
<td>global campus</td>
<td></td>
<td></td>
</tr>
<tr>
<td>49996</td>
<td>pbc</td>
<td>Personal Business Commitments</td>
<td>60</td>
</tr>
<tr>
<td>46199</td>
<td>bond</td>
<td>Buy On Demand</td>
<td>13</td>
</tr>
<tr>
<td>44582</td>
<td>human resources</td>
<td></td>
<td></td>
</tr>
<tr>
<td>30806</td>
<td>global print</td>
<td></td>
<td></td>
</tr>
<tr>
<td>29698</td>
<td>eamt</td>
<td>Electronic Asset Management Tool</td>
<td>1</td>
</tr>
<tr>
<td>25921</td>
<td>travel</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20940</td>
<td>jobs</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Observations**

- All top 10 queries are navigational queries
- 5 of the top 10 queries are acronyms
- 7 of the top 10 are one word queries and 3 are two word queries
Another example

1. Total number of hits on the intranet = 7657
2. The acronym PBC has 3 different expansions
3. Top required hit refers to Personal Business Commitment
This query is a little bit more complex

IBM Italy

IBM Spain

You and IBM - Italia

Personal Business Commitments

Il successo della IBM dipende da come ciascuno di noi obiettivi strategici della Compagnia. Il programma PBC fissano i propri obiettivi per l'anno, ricevono feedback valutati sulla base delle loro performance.

Ad inizio anno, viene richiesto ad ogni singolo dipendente un obiettivo ed i valori IBM. Questi obiettivi devono guidare l'obiettivo nell'ambito della professione, la prestazione complessiva. Questa valutazione sarà futuri incrementi salariali e opportunità di carriera.

L'intera struttura del programma PBC è disegnato per contributo relativo dato da ciascun dipendente alla un ambiente in cui ogni dipendente sia stimolato a premiare i maggiori contribuenti, motivare quella sopraidentificare i "low performer" che devono migliorare compagnia ricevano i maggiore riconoscimenti. Le fasi del processo annuale della PBC sono:

You and IBM - España

Personal Business Commitments

El éxito de IBM depende de cada uno de nosotros. De nuestra contribución a la dirección estratégica de IBM, a nivel mundial, establezcan sus objetivos anual, y son evaluados por su contribución.

Al comienzo de cada año, cada empleado debe fijar objetivos de negocio globales de la compañía. Estos director evaluarán los resultados, en comparación con el rendimiento. Esta calificación constituirá un factor para incrementos salariales y oportunidades de carrera.

El PBC está diseñado para reforzar la cultura de alto basándose en su contribución a la compañía. Para un clima en el que cada empleado siente la necesidad a los que más contribuyen, empujar a otros para que empleados que más aportan a la compañía reciban la.

Existen tres pasos en el proceso anual del PBC. Siga usted y lo que usted pueda esperar de ella.

Paso 1: Establecer Sus Objetivos
Difficulties captured in Fagin et al 2003

- **Axiom 1.** Intranet documents are often created for simple dissemination of information, rather than to attract and hold the attention of any specific group of users.

- Describes navigational queries and highlights the difficulty in identifying such pages on the intranet.

- **Axiom 3.** Intranets are essentially spam-free.

- Refers to company-specific queries such as employees, product-names etc. and the difficulties encountered in a geographically disperse organization.
Existing System

- Link-analysis + conventional IR-ranking heuristics
- The correct result for both “idp” and “pbc” not in the top 50 results

Hard to tune generic ranking algorithms to consistently provide top results for all navigational queries!
Answering Navigational Queries

- **Option 1: Do nothing special**
  - Trust generic content / link-based ranking
    1. *Unfortunately, this does not work (existing system)*
    2. *Very hard for a generic ranking algorithm to consistently identify the top results for every navigational query*

  *Can we special-case?*

- **Option 2: Special-case navigational queries**
  - Assume we can identify the top result(s) for each navigational query
  - Maintain a separate navigational index over just these result pages
  - Unfortunately, not so simple!!
    1. *A priori we don't know all the queries!!*
    2. *(Even if we did, we would need to classify queries as being navigational or otherwise)*

  *Instead of queries can we start from documents?*
Option 3

- Run through the full collection and identify navigational pages
- With each identified page associate appropriate query terms
- Separately create a navigational index using these terms and identified pages
Outline for remainder …

• Offline Processing
  – Identify navigational pages
    • Local Analysis
    • Global Analysis
  – Associate appropriate terms with each navigational page
    • Term variant generation

• Runtime
  – Almost like a database lookup of query term(s) *(a little bit more !)*
Local analysis

- Per-page analysis to identify candidate navigational pages
- Outline
  - Extract navigational features <Title, URL, etc.>
  - Pattern and dictionary matches
    - \A\W+(.*)'s <Home>\b
    - \b<Home> of (.*)<junk>\Z
  .
  .
  .

Output = Candidate Navigational page (Y / N). If Y then feature value associated with page

Local Analysis Template
Different Local analyses

IBM Global Services Security Home

G J Chaitin Home Page

Title Extraction

Matching title patterns

Titles

Title Name

IBM Global Services Security

Title Extraction

Matching title patterns

Titles

Dictionary Match

Person name dictionary

Person name in title

G J Chaitin

URL Extraction

Matching URL patterns

URLs

URL Name

1. marketing
2. isc
3. chis

NavPanel Extraction

Self link identification

NavPanels

LA

LA

LA

Person name dictionary = employee directory

1. http://w3-03.ibm.com/marketing/
Why Global Analysis?

- We have individual pages and associated feature values e.g.,

  G J Chaitin → G J Chaitin’s home page

Can we put these directly into the index!!

Unfortunately not so simple!!
This website contains Greek letters and other mathematical symbols. If "Ω" isn't a capital Greek letter Omega, you should switch to another browser, for example, MS IE or Mozilla Firefox.

This website contains most of Chaitin's published papers, many book chapters, and the LISP, Java, C, and Mathematica software for Chaitin's Springer-Verlag trilogy. It also contains interviews and reviews of Chaitin's books.

Contents

- Latest News
- Recent Books
- LISP Applet
- Books with LISP Software
- Collections of Interviews
- Essays on Leibniz & Other Popular Articles
- Photo Album & Time-Line
- Brief Biography
- Latest Publications
- Complete List of Publications
IBM Club presents Gaslight Night

Sinbad, October 14

Tickets for the Gaslight Theatre’s production, “Sinbad,” for October 14 are now on sale at the IBM Club Box Office. The show will start at 7:00 p.m. (doors open at 6:15 p.m.) on Friday evening.

Tickets for reserved seats are $8.00 each ($14.95 public price) and are sold by tables on a first-come-first-served basis.

More about “Sinbad”

In a word or two—comes a man we’ve never seen nor heard of... uproot a swashbuckling, gun-slinging, musical sort taken directly from the pages of “1001 Arabian Nights.” Join Sinbad on the most terrifying and romantic voyage of his life. Cheer as our hero pursues the greatest prize known to man, The Treasure of Alexander the Great.

Hidden on the Mythical Island of Colossus, the treasure is protected by legendary beasts conjured up from the land beyond! 8000 at our villain, the evil Emir of Dhabul and his lovely assistant, Suliana Balak, try to thwart Sinbad at every turn. Laugh at the zany antics of the court magician, Gojo the Great. Enjoy the tender moments shared by Sinbad and the love of his life, the beautiful Princess Shalimar. Don’t be left on shore!

Contact Kathy Carlisle at ext 2209 for more details.
What we do in Global Analysis

- Site Root Analysis
  - Personal Home Pages
  - Named Title Home Pages
  - Named URL Home Pages

- Extended Site Root Analysis
  - Rooted Home Pages

- Anchor Text Analysis
Site root analysis template

- **Grouping Input**
  - performed on feature values of a particular local analysis (e.g., G J Chaitin)
- **Grouping function**
  - Equality grouping function (more complex later)
- **Forest Generation**
  - URL provides clues about navigational pages
- **Marker inputs**
  - Other local analysis may provide more evidence and therefore used as markers
Specific instances of site root analysis

- Intranet pages
- Grouping Input
- Site Root Analysis
- NamedTitle home pages
- Exact match on feature values from LA

NavPanel Self-link Marker Input
URL Forest generation

All pages have the same Title feature-value !! (e.g., IBM Tucson)
Apply Marker Inputs

- host1/b/a
  - host1/b/a/~/user1/
    - host1/b/a/~/user1/pub
  - host1/b/a/x_index.htm/
- host1/b/c
  - host1/b/c/d
    - host1/b/c/d/e/index.html
  - host1/b/c/home.html
  - host1/b/c/d/e/index.html?a=us
  - host1/b/c/d/e/index.html?a=uk
Retain all roots and output navigational pages
Extended Site Root

• Highly restrictive local analysis
• But what about the remaining pages?
• Observation
  – Page titles exhibit hierarchies
Use the same trick as before ………..

Marker Inputs are output of other GA results

Grouping Input = Title

Hierarchical matching on titles

Hierarchical Grouping function

Intranet pages

GA_p

GA_t

GA_u

LA_n

Site Root Analysis

Rooted home pages
Anchor text


Focus on precision enforced by restricted anchor-text extraction
The most frequent “anchor text” associated with page as feature value
Term Variant Generation
Why term variant generation?

- User queries may not exactly match feature-values
  - People names
    - Ching-Tien T. (Howard) Ho
    - User queries include <Ho Ching-Tien> <Tien Ho>, <Ho, Howard>
  - Acronym expansions
    - gts → Global Technology Services
  - N-gram variant generation
    - Title: reimbursement of travel expenses
    - reimbursement, travel expenses, reimbursement travel, reimbursement of travel, reimbursement expenses
Recapture what we did

• Offline Processing
  – Identify navigational pages
    • Local Analysis
    • Global Analysis
  – Associate appropriate terms with each navigational page
    • Term variant generation
  – Output is 4 semantic buckets
    • Personal, NamedTitle, NamedURL and Rooted
    • For analysis NamedTitle, NamedURL and Rooted are split into Acronym vs Non-Acronym
Experiments
Experimental Setup

- Crawled pages: 5.5 m (crawled between Jun-Aug 2006)
- 55,000 pages retained after local and global analysis
- Gold standard: 345 queries, 446 pages.
- Queries for which our system provides an answer: 321
## Performance by semantic buckets

<table>
<thead>
<tr>
<th>MRR</th>
<th>all</th>
<th>1kw</th>
<th>2kw</th>
<th>3kw</th>
<th>&gt;3kw</th>
</tr>
</thead>
<tbody>
<tr>
<td>Personal</td>
<td>0.59</td>
<td>0.09</td>
<td>0.92</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>A-NamedTitle</td>
<td>0.36</td>
<td>0.38</td>
<td>0</td>
<td>0.38</td>
<td>0</td>
</tr>
<tr>
<td>NamedTitle</td>
<td>0.20</td>
<td>0.14</td>
<td>0.18</td>
<td>0.62</td>
<td>1</td>
</tr>
<tr>
<td>A-NamedURL</td>
<td>0.27</td>
<td>0.27</td>
<td>0</td>
<td>0.3</td>
<td>0.5</td>
</tr>
<tr>
<td>NamedURL</td>
<td>0.24</td>
<td>0.25</td>
<td>0.15</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>A-Rooted</td>
<td>0.13</td>
<td>0.16</td>
<td>0</td>
<td>0.06</td>
<td>0</td>
</tr>
<tr>
<td>Rooted</td>
<td>0.14</td>
<td>0.09</td>
<td>0.16</td>
<td>0.35</td>
<td>1</td>
</tr>
</tbody>
</table>

- Personal bucket in general does well
- For NamedTitle and NamedURL acronym dictionaries are useful
Ranking within a semantic bucket has little effect

- Ranking within semantic bucket is a black-box (Lucene)
- Randomization of the ranks within a semantic bucket marginally changes the overall performance

<table>
<thead>
<tr>
<th></th>
<th>Personal</th>
<th>A-Named Title</th>
<th>Named Title</th>
<th>A-Named URL</th>
<th>Named URL</th>
<th>A-Rooted</th>
<th>Rooted</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-run</td>
<td>0.59</td>
<td>0.36</td>
<td>0.20</td>
<td>0.27</td>
<td>0.24</td>
<td>0.13</td>
<td>0.14</td>
</tr>
<tr>
<td>Rand</td>
<td>0.60</td>
<td>0.33</td>
<td>0.18</td>
<td>0.27</td>
<td>0.22</td>
<td>0.13</td>
<td>0.10</td>
</tr>
</tbody>
</table>

Quality of identified navigational pages very high
Ranking across buckets matters

- Order 1: Personal, Named Title, URLName
- Order 2: Personal, URLName, Named Title
- Order 3: Personal, Acronym, Non-Acronym (Best order based on result MRR)
- Order 4: Simple Statistical Model
- Order 5: Adversarial (Opposite ranking of Order 3)
- Current System (W3)

<table>
<thead>
<tr>
<th>Order</th>
<th>Overall MRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Order 1</td>
<td>0.32</td>
</tr>
<tr>
<td>Order 2</td>
<td>0.32</td>
</tr>
<tr>
<td>Order 3</td>
<td>0.33</td>
</tr>
<tr>
<td>Order 4</td>
<td>0.34</td>
</tr>
<tr>
<td>Order 5</td>
<td>0.21</td>
</tr>
<tr>
<td>Current System</td>
<td>0.18</td>
</tr>
</tbody>
</table>

Adversarial ranking is better than the existing system!
Intranet Local Search

IBM Italy

IBM Spain

You and IBM - Italia

You and IBM - España

Personal Business Commitments

Personal Business Commitments

PBC

El éxito de IBM depende de cada uno de nosotros. D de nuestra contribución a la dirección estratégica de IBM, a nivel mundial, establezcan sus objetivos anual, y son evaluados por su contribución.

Al comienzo de cada año, cada empleado debe fijar objetivos de negocio globales de la compañía. Estos directores evaluarán sus resultados, en comparación con rendimiento. Esta calificación constituirá un factor para incrementos salariales y oportunidades de carrera.

El PBC está diseñado para reforzar la cultura de alto basándose en su contribución a la compañía. Para que un clima en el que cada empleado sienta la necesidad a los que más contribuyen, empújar a otros para que empleados que más aporten a la compañía reciban las.

Existen tres pasos en el proceso anual del PBC. Siga usted y lo que usted pueda esperar de ella.

Paso 1: Establecer Sus Objetivos
Not all queries are geo-sensitive
Intranet Local Search Challenges

- **Offline**
  - Geo-identification of pages
  - Site-specific, country-specific, region-specific dictionaries

- **Runtime**
  - Detect geo-sensitive queries (Future work)
Intranet Local Search

- Assuming geo-sensitive queries have been identified perfectly

<table>
<thead>
<tr>
<th>Country</th>
<th>Region</th>
<th>MRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>UK</td>
<td>EMEA</td>
<td>0.4858</td>
</tr>
<tr>
<td>DE</td>
<td>EMEA</td>
<td>0.4879</td>
</tr>
<tr>
<td>CA</td>
<td>Americas</td>
<td>0.4433</td>
</tr>
<tr>
<td>JP</td>
<td>Asia-Pacific</td>
<td>0.484</td>
</tr>
<tr>
<td>AU</td>
<td>Asia-Pacific</td>
<td>0.493</td>
</tr>
</tbody>
</table>
Conclusions

- *A priori* offline identification of navigational pages provides room for improving MRR.
- Stratification is a big win: semantic buckets, geography, query length, etc.
- Domain dictionaries have significant values.
- Ranking among the buckets is more useful than ranking within the buckets.
Marking and pruning example
Example geography statistics

- **NamedURL:**
  - Sites: N/A:101 ZRL:2
  - Countries: CH:2 JP:8 CA:2 US:15 N/A:76
  - Regions: AMERICAS:63 AP:8 EMEA:11 N/A:19 LATINAMERICA:2

- **NamedTitle:**
  - Sites: N/A:18
  - Countries: CA:1 JP:1 US:9 N/A:7
  - Regions: AMERICAS:10 AP:1 EMEA:1 N/A:6

- **Rooted:**
  - Sites: HURSLEY:2 N/A:728 HRL:7 ARC:1
  - Regions: AMERICAS:61 AP:26 EMEA:471 N/A:164 LATINAMERICA:16
Ranking among types is more important than in-type ranking

- Ranking within semantic bucket is provided by Lucene.
- Randomization of the ranks within a semantic bucket does not change the overall performance.

<table>
<thead>
<tr>
<th>MRR</th>
<th>all</th>
<th>1kw</th>
<th>2kw</th>
<th>3kw</th>
<th>&gt;3kw</th>
</tr>
</thead>
<tbody>
<tr>
<td>Personal</td>
<td>0.60</td>
<td>0.09</td>
<td>0.93</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>A-NamedTitle</td>
<td>0.33</td>
<td>0.38</td>
<td>0</td>
<td>0.30</td>
<td>0</td>
</tr>
<tr>
<td>NamedTitle</td>
<td>0.18</td>
<td>0.12</td>
<td>0.16</td>
<td>0.61</td>
<td>1</td>
</tr>
<tr>
<td>A-NamedURL</td>
<td>0.27</td>
<td>0.26</td>
<td>0</td>
<td>0.28</td>
<td>0.61</td>
</tr>
<tr>
<td>NamedURL</td>
<td>0.22</td>
<td>0.23</td>
<td>0.17</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>A-Rooted</td>
<td>0.13</td>
<td>0.17</td>
<td>0</td>
<td>0.03</td>
<td>0</td>
</tr>
<tr>
<td>Rooted</td>
<td>0.10</td>
<td>0.05</td>
<td>0.14</td>
<td>0.31</td>
<td>0.61</td>
</tr>
</tbody>
</table>
All these pages have the exact same title IBM Tucson Home
This website contains Greek letters and other mathematical symbols. If "Ω" isn't a capital Greek letter Omega, you should switch to another browser, for example, MS IE or Mozilla Firefox.

This website contains most of Chaitin's published papers, many book chapters, and the LISP, Java, C, and Mathematica software for Chaitin's Springer-Verlag trilogy. It also contains interviews and reviews of Chaitin's books.

Contents
- Latest News  - Recent Books  - LISP Applet  
- Books with LISP Software  - Collections of Interviews  
- Essays on Leibniz & Other Popular Articles  
- Photo Album & Time-Line  - Brief Biography  
- Latest Publications  - Complete List of Publications
Sample navigational panel