Navigating the Intranet with High Precision

Huaiyu Zhu
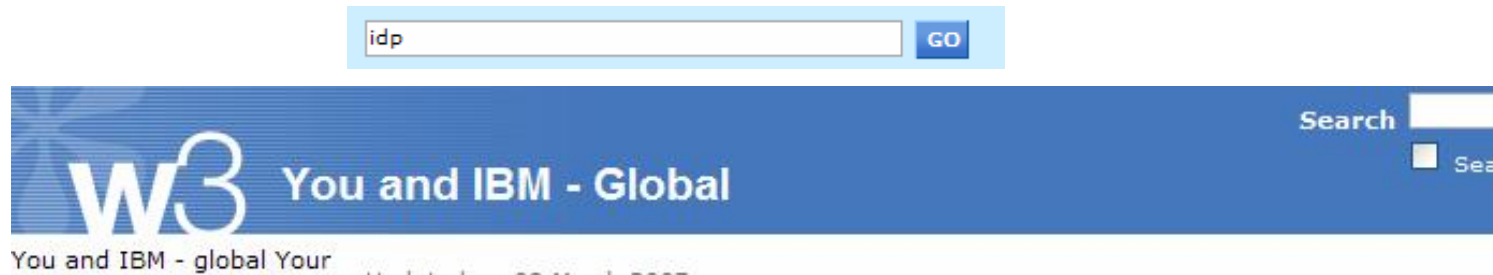Alexander Löser
Sriram Raghavan
Shivakumar Vaithyanathan

IBM Almaden Research Center

What is this paper about ?

idp **GO**

W3 You and IBM - Global

Search

You and IBM - global Your

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

progress and using the plan as a living document.

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

- **Valued skills :** Access the Market Valued Skills quickview to learn about the skills m

- **Find out more :** Descriptions of new IDP features and Frequently Asked Questions (

Terms of use

Top unique 10 queries on IBM Intranet

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

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

pbc | GO

You and IBM - Global

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

- IBM strategy
- IBM values
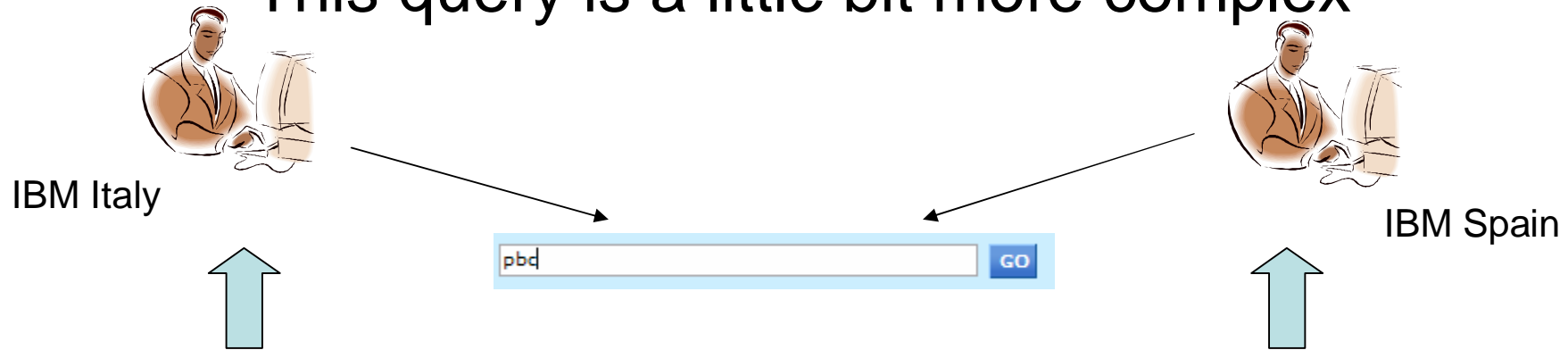- Individual Development Plan

Site secured by
IBM intranet password

Is this your first time here this year?
Create your new PBC here

Work with my PBC
Already started?
Work on your draft PBCs here

To protect the privacy of your information, you must shut down all windows so that you will be logged off the IBM Intranet Password S

Terms of use

This query is a little bit more complex

IBM Italy

pbc [GO]

IBM Spain

W3 You and IBM - Italia

You and IBM - Italia
Manager portal
Your money
Your health
Your life
Your career
Ask HR

You and IBM - Europe

Ask HR

Reviewed on 11/23/2006

Set Your Goals >

Personal Business Commitments

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

Ad inizio anno, viene richiesto ad ogni singolo dipend
obiettivi ed i valori IBM. Questi obiettivi devono guid
raggiungimento degli obiettivi nell'ambito della profes
la prestazione complessiva. Questa valutazione sarà
futuri incrementi salariali e opportunità di carriera.

L'intera struttura del programma PBC è disegnata pe
contributo relativo dato da ciascun dipendente alla o
un ambiente in cui ogni dipendente sia stimolato a o
premiare i maggiori contributori, motivare quelli sopra
indentificare i "low performer" che devono migliorare
compagnia ricevono i maggiori riconoscimenti.
Le fasi del processo annuale della PBC sono:

W3 You and IBM - España

You and IBM SA - España
You and IBM GSE SA - España
Manager portal
About you

Alphabetical index
  IBM SA | IBM GSE SA

Ask HR

Reviewed on 01/05/2007

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 anua
formación, y son evaluados por su contribución.

Al comienzo de cada año, cada empleado debe fijar
objetivos de negocio globales de la compañía. Estos
director evaluará sus resultados, en comparación co
rendimiento. Esta calificación constituirá un factor c
para incrementos salariales y oportunidades de carre

El PBC está diseñado para reforzar la cultura de alto
basándose en su contribución a la compañía. Para p
un clima en el que cada empleado sienta la necesida
a los que más contribuyen, empujar a otros para que
empleados que más aportan a la compañía reciben 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*
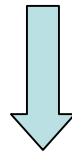
Can we special-case ?

- **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)*
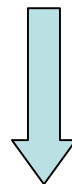
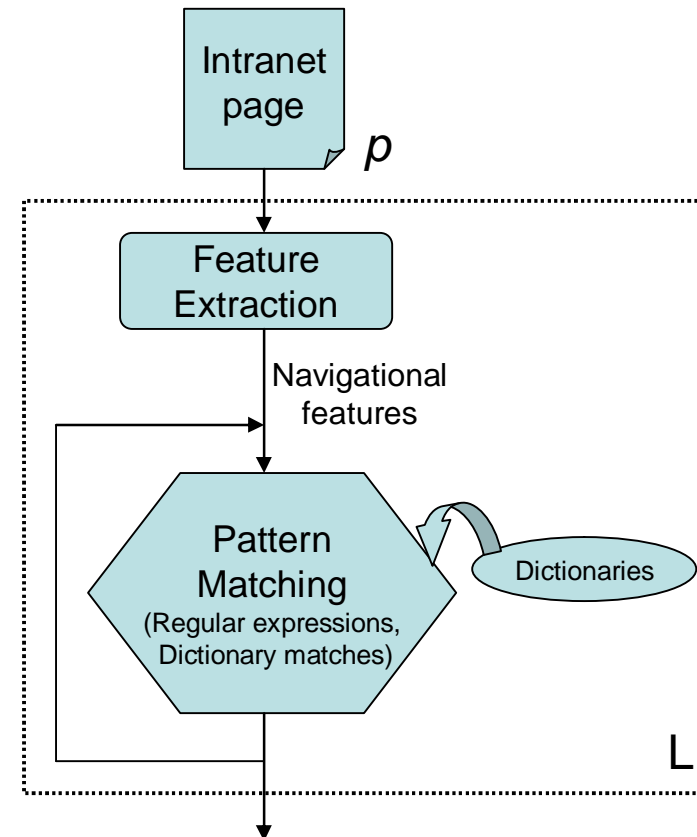*Instead of queries can we start from documents ?*

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 !)
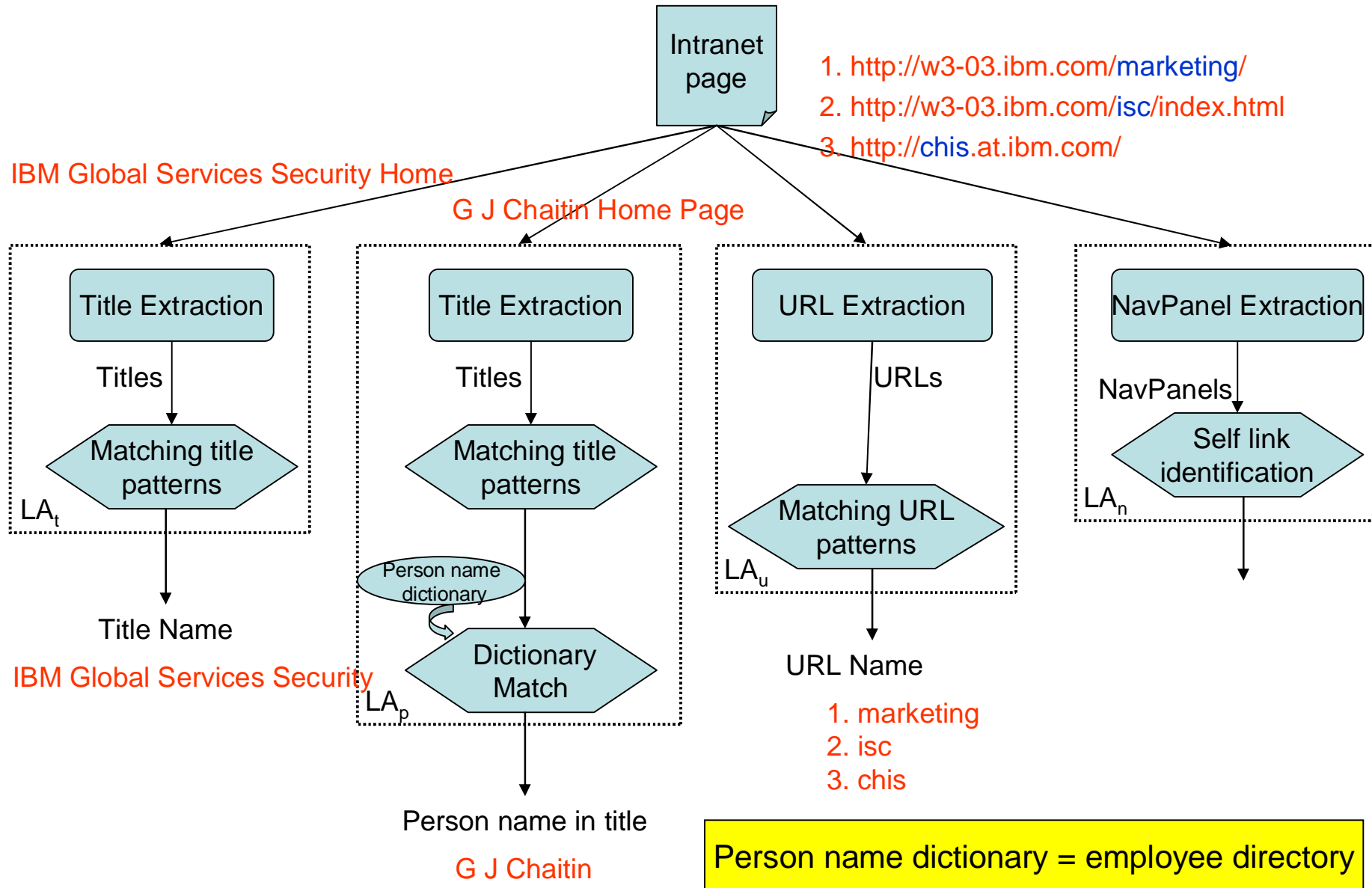
Local analysis

- 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
    - .
    - .
    - .
    - .

Intranet page

p

Feature Extraction

Navigational features

Pattern Matching
(Regular expressions, Dictionary matches)

Dictionaries

L

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

Local Analysis Template

Different Local analyses

Intranet page

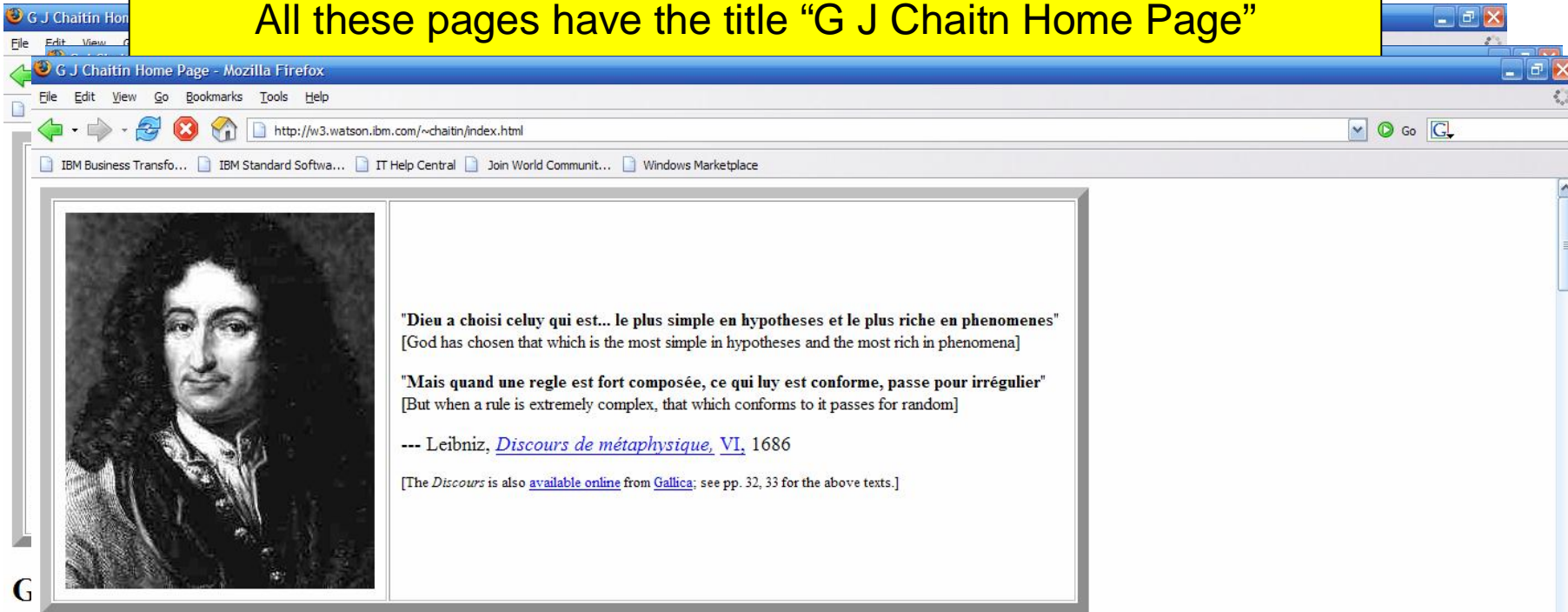1. http://w3-03.ibm.com/marketing/
2. http://w3-03.ibm.com/isc/index.html
3. http://chis.at.ibm.com/

IBM Global Services Security Home

G J Chaitin Home Page

Title Extraction

Titles

Matching title patterns

LA_t

Title Name

IBM Global Services Security

Title Extraction

Titles

Matching title patterns

Person name dictionary

Dictionary Match

LA_p

Person name in title

G J Chaitin

URL Extraction

URLs

Matching URL patterns

LA_u

URL Name

1. marketing
2. isc
3. chis

NavPanel Extraction

NavPanels

Self link identification

LA_n

Person name dictionary = employee directory

Why Global Analysis ?

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

   G J Chaitin $\rightarrow$ G J Chaitin's home page

   Can we put these directly into the index !!
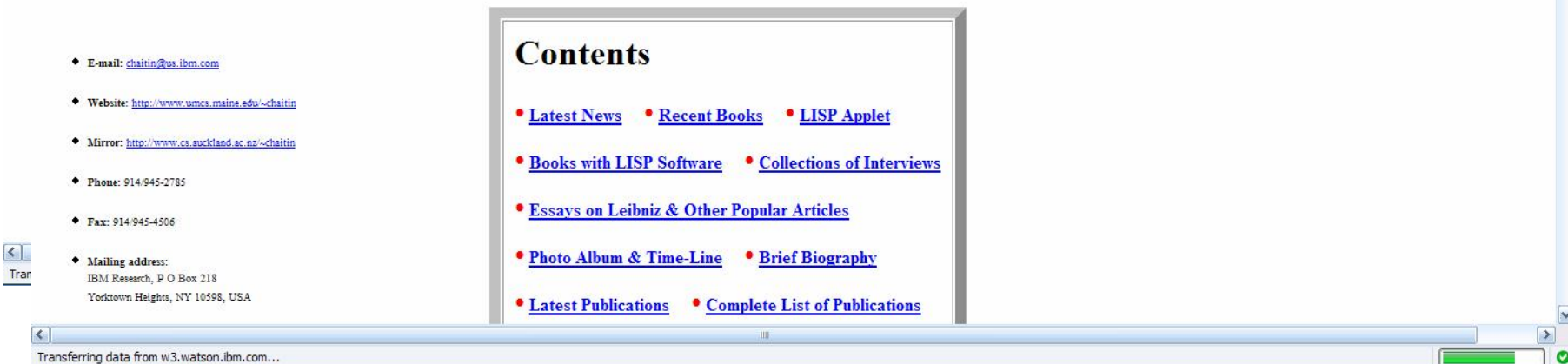
   Unfortunately not so simple !!

All these pages have the title "G J Chaitn Home Page"

G J Chaitin Home Page - Mozilla Firefox

File   Edit   View   Go   Bookmarks   Tools   Help

http://w3.watson.ibm.com/~chaitin/index.html

IBM Business Transfo...   IBM Standard Softwa...   IT Help Central   Join World Communit...   Windows Marketplace

"Dieu a choisi celuy qui est... le plus simple en hypotheses et le plus riche en phenomenes"
[God has chosen that which is the most simple in hypotheses and the most rich in phenomena]

"Mais quand une regle est fort composée, ce qui luy est conforme, passe pour irrégulier"
[But when a rule is extremely complex, that which conforms to it passes for random]

--- Leibniz, *Discours de métaphysique,* VI, 1686

[The *Discours* is also available online from Gallica; see pp. 32, 33 for the above texts.]

G J Chaitin Home Page

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.

- E-mail: chaitin@us.ibm.com
- Website: http://www.umcs.maine.edu/~chaitin
- Mirror: http://www.cs.auckland.ac.nz/~chaitin
- Phone: 914/945-2785
- Fax: 914/945-4506
- Mailing address:
  IBM Research, P O Box 218
  Yorktown Heights, NY 10598, USA

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**

Transferring data from w3.watson.ibm.com...

All these pages have the exact same title "IBM Tucson | Tucson Home"

IBM Tucson | Tucson Home - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://w3.tucson.ibm.com/docs/stories/past/2005/events/05_08_30_gaslight.html

IBM Business Transfo... IBM Standard Softwa... IT Help Central Join World Communit... Windows Marketplace

Search w3   GO
        Search IBM Tucson

w3 IBM Tucson

w3 Home | BluePages | HelpNow | Feedback

Tucson Home

STG Home
IBM Home
Site Calendar
Site Links
Site Communications
People
Visitor Information
Story Archives
Protecting You and IBM

May 9, 2007

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.

THE GASLIGHT THEATRE

More about "Sinbad"

In a world of heroes comes a man we've all been waiting for ... Sinbad! A swashbuckling, high seas, musical spoof 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, this treasure is protected by legendary beasts conjured up from the land beyond! Boo as our villains, the evil Emir of Dhaibul and his lovely assistant, Sultana Dalilah, try to thwart Sinbad at every turn. Laugh at the zany antics of the court magician, Golgo 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.

Terms of use

Related Links

Site Calendar
[IBM Tucson]
Get IBM news here.
Get customized news by signing in to On Demand Workplace.

Tucson Information

Select One...

Congratulate or Thank a colleague ...send an e-card!
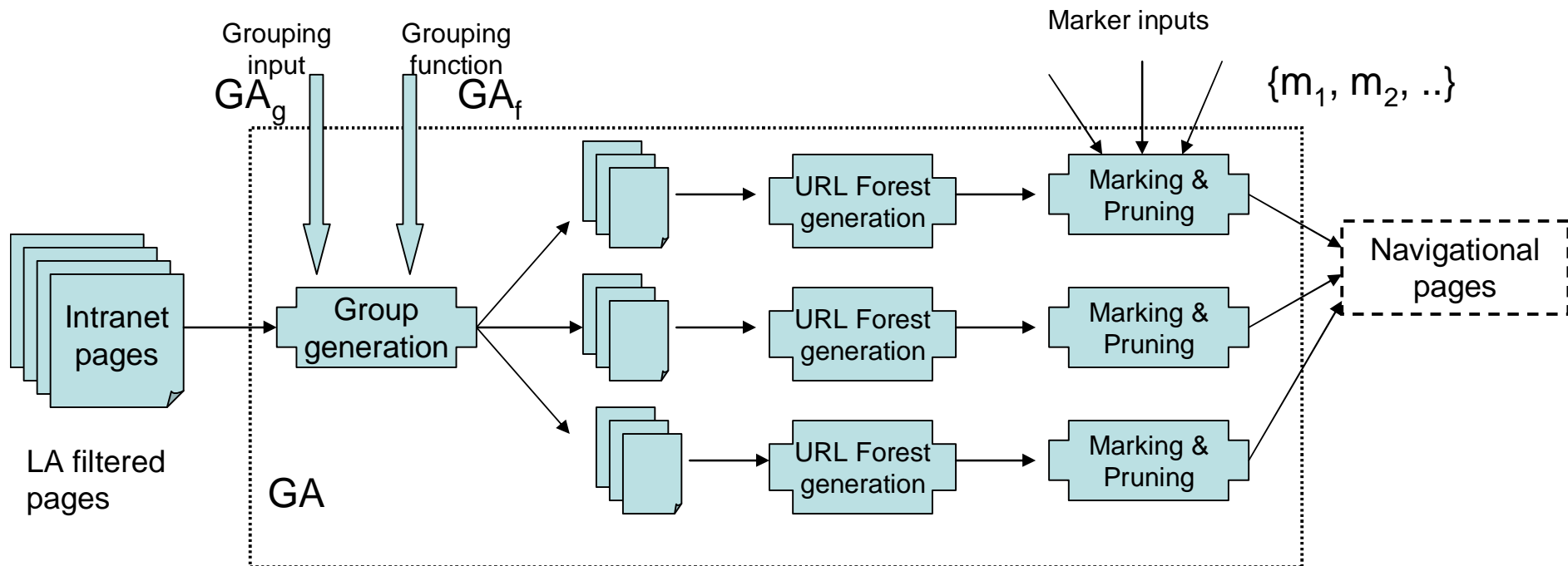
IBM Stock Quote

Done

Find: ibm tucson   Find Next Find Previous Highlight all Match case
Done

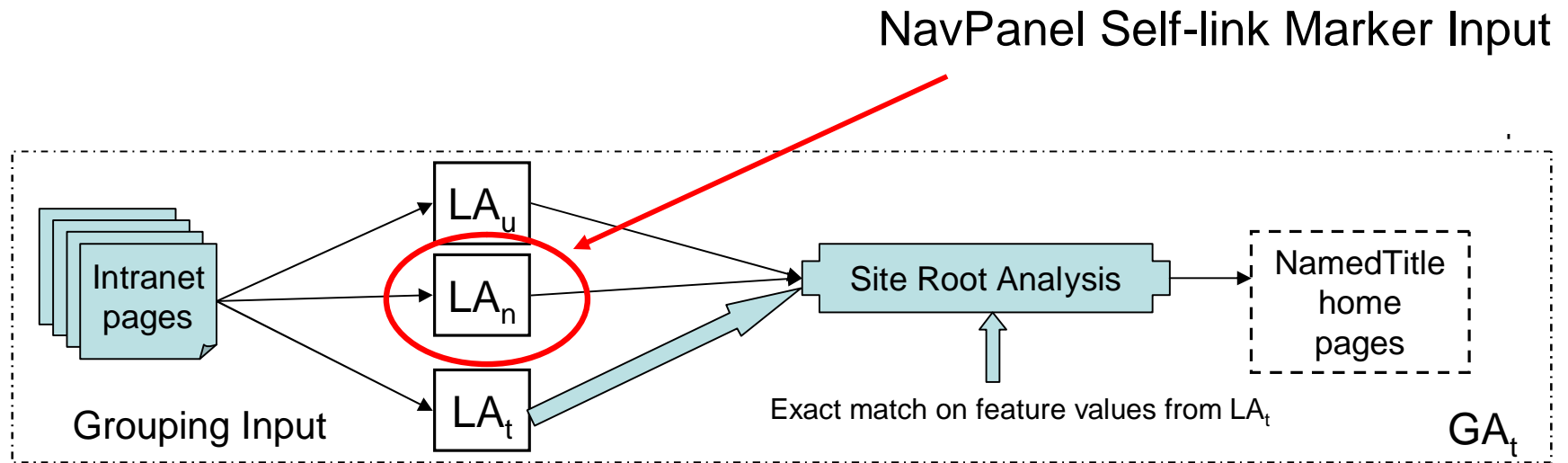# What we do in Global Analysis

- 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

NavPanel Self-link Marker Input

Intranet pages

LA_u

LA_n

LA_t

Grouping Input

Site Root Analysis

Exact match on feature values from LA_t

NamedTitle home pages

GA_t

URL Forest generation

All pages have the same Title feature-value !! (e.g., IBM Tucson)
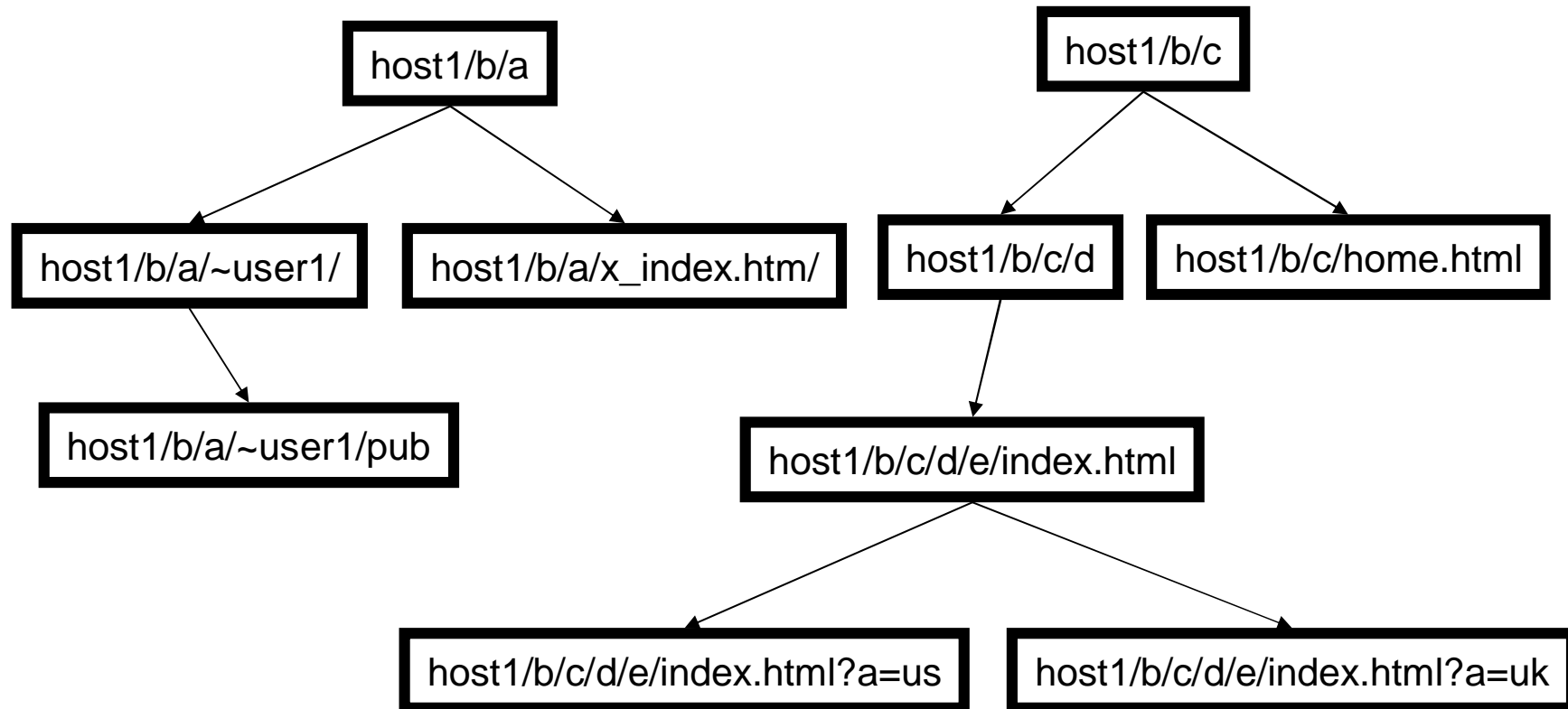
Apply Marker Inputs
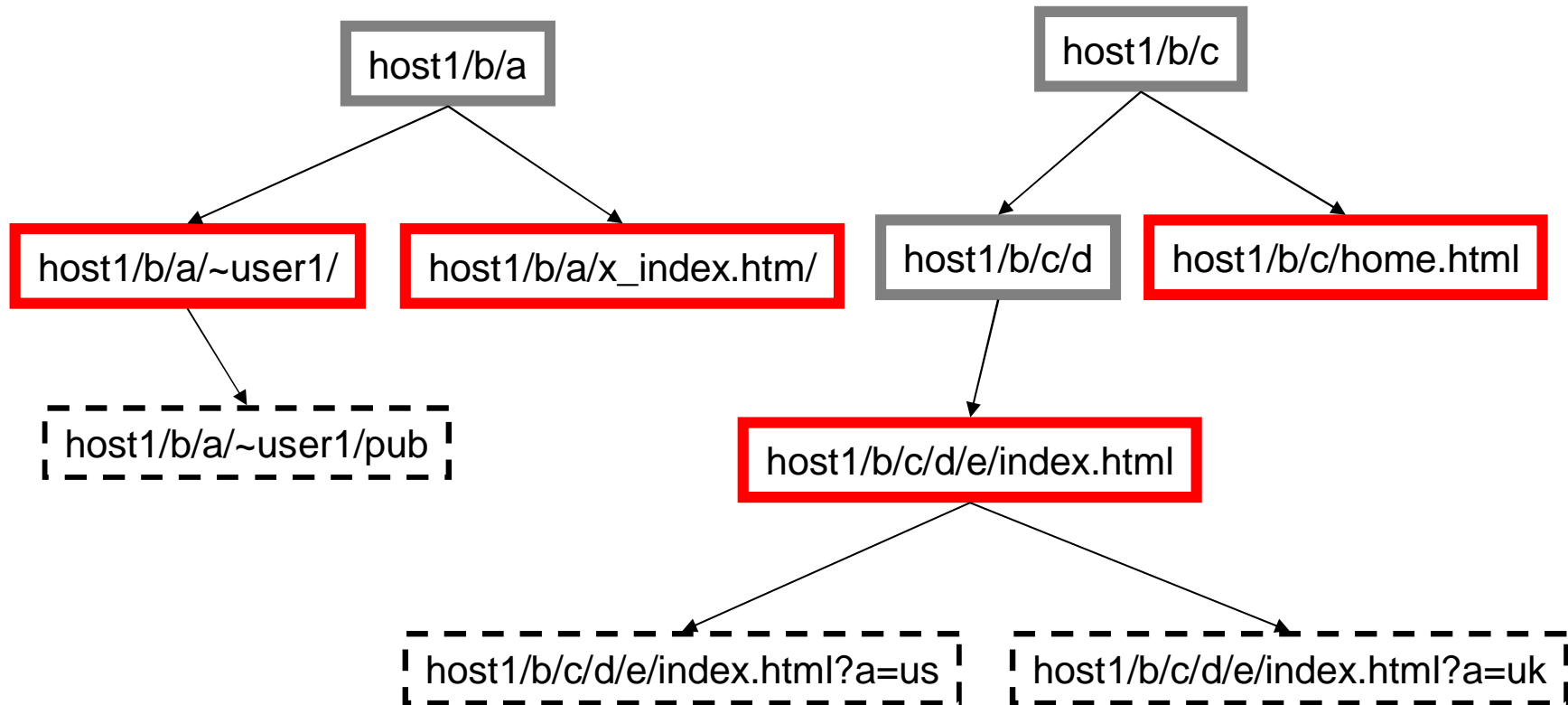
Retain all roots and output navigational pages
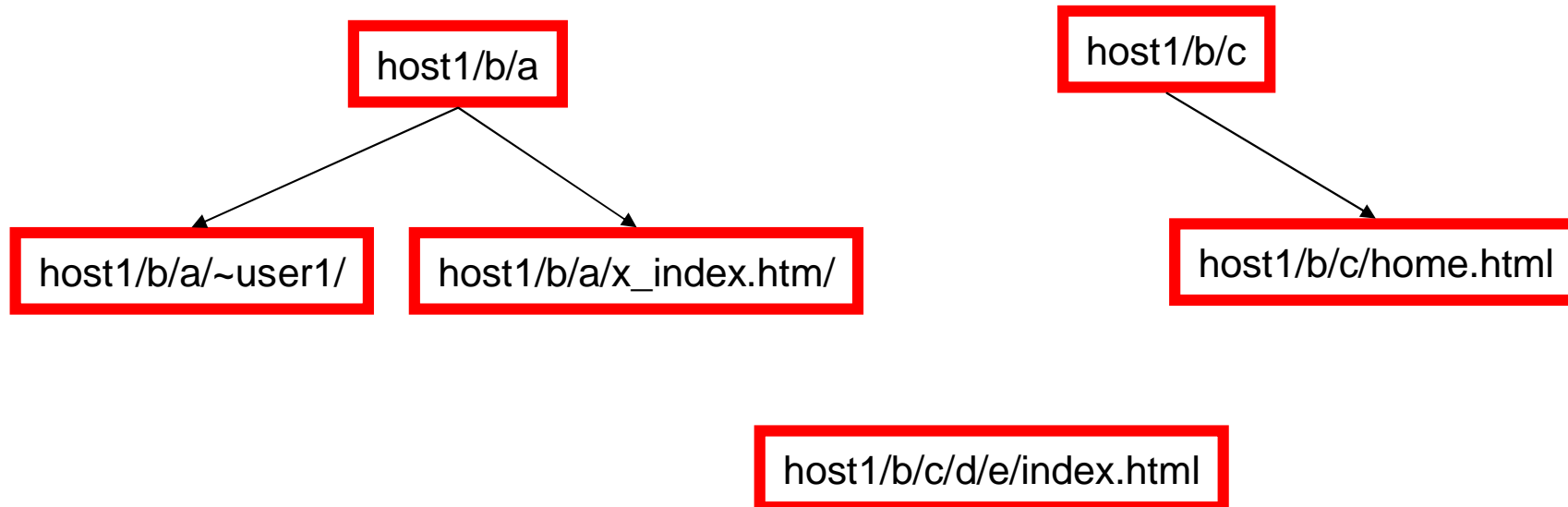
Extended Site Root

- Highly restrictive local analysis
- But what about the remaining pages ?
- Observation
  - Page titles exhibit hierarchies

....
Software | Business transformation | IT Infrastructure Portal | Usage Metering and Billing
Software | Business transformation | IT Infrastructure Portal | Usage Metering and Billing Rules
Software | Business transformation | IT Infrastructure Portal | Welcome to the IT Rep page

Software | Business transformation | IT Security | Application and Information Compliance
Software | Business transformation | IT Security | General Security Information

.

.

Use the same trick as before ....

Marker Inputs are output of other GA results

GA_p

GA_t

GA_u

LA_n

Intranet pages

Title Extraction

Site Root Analysis

Rooted home pages

Hierarchical matching on titles

Grouping Input = Title

Hierarchical Grouping function

Anchor text

- http://rossr.torolab.ibm.com:80/
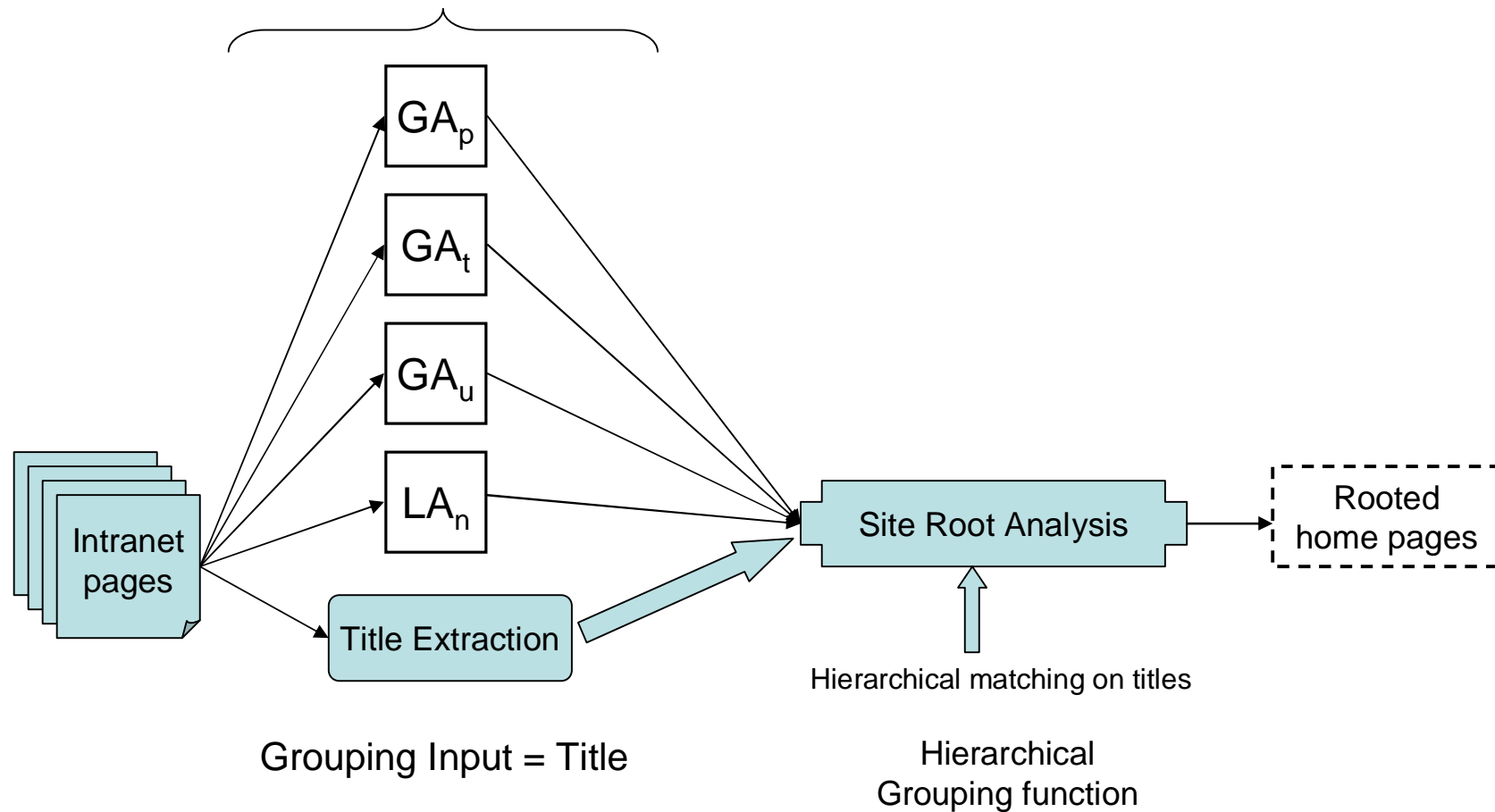
ID related links
- DB2 Technical Support Library
- DB2 ID Intranet homepage
- Opening documentation defects
- DB2 ID Internal Project Portal (authentication req'd)
- DB2 ID personnel database
- DB2 UDB Mission Control
- David Sky's DB2 Information Architecture site
- IDWB homepage and Eclipse documentation
- Authoring Tools Information Center

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

MRR	all	1kw	2kw	3kw	>3kw
Personal	0.59	0.09	0.92	1	1
A-NamedTitle	0.36	0.38	0	0.38	0
NamedTitle	0.20	0.14	0.18	0.62	1
A-NamedURL	0.27	0.27	0	0.3	0.5
NamedURL	0.24	0.25	0.15	-	-
A-Rooted	0.13	0.16	0	0.06	0
Rooted	0.14	0.09	0.16	0.35	1

- 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

	Personal	A-Named Title	Named Title	A-Named URL	Named URL	A-Rooted	Rooted
1-run	0.59	0.36	0.20	0.27	0.24	0.13	0.14
Rand	0.60	0.33	0.18	0.27	0.22	0.13	0.10

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)

Order	Overall MRR
Order 1	0.32
Order 2	0.32
Order 3	0.33
Order 4	0.34
Order 5	0.21
Current System	0.18

Adversarial ranking is better than the existing system !

Intranet Local Search

IBM Italy

IBM Spain

pbc   GO

w3 You and IBM - Italia

You and IBM - Italia

Manager portal

Your money

Your health

Your life

Your career

Ask HR

You and IBM - Europe

Ask HR

Reviewed on 11/23/2006

Set Your Goals >

Personal Business Commitments

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

Ad inizio anno, viene richiesto ad ogni singolo dipend
obiettivi ed i valori IBM. Questi obiettivi devono guid
raggiungimento degli obiettivi nell'ambito della profes
la prestazione complessiva. Questa valutazione sarà
futuri incrementi salariali e opportunità di carriera.

L'intera struttura del programma PBC è disegnata pe
contributo relativo dato da ciascun dipendente alla c
un ambiente in cui ogni dipendente sia stimolato a o
premiare i maggiori contributori, motivare quelli sopra
indentificare i "low performer" che devono migliorare
compagnia ricevono i maggiori riconoscimenti.
Le fasi del processo annuale della PBC sono:

w3 You and IBM - España

You and IBM SA - España

You and IBM GSE SA - España

Manager portal

About you

Alphabetical index

IBM SA | IBM GSE SA

Ask HR

Reviewed on 01/05/2007

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 anua
formación, y son evaluados por su contribución.

Al comienzo de cada año, cada empleado debe fijar
objetivos de negocio globales de la compañía. Estos
director evaluará sus resultados, en comparación co
rendimiento. Esta calificación constituirá un factor c
para incrementos salariales y oportunidades de carre

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

Not all queries are geo-sensitive

IBM Italy

IBM Spain

Search for: issi **GO**

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

Country	Region	MRR
UK	EMEA	0.4858
DE	EMEA	0.4879
CA	Americas	0.4433
JP	Asia-Pacific	0.484
AU	Asia-Pacific	0.493

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
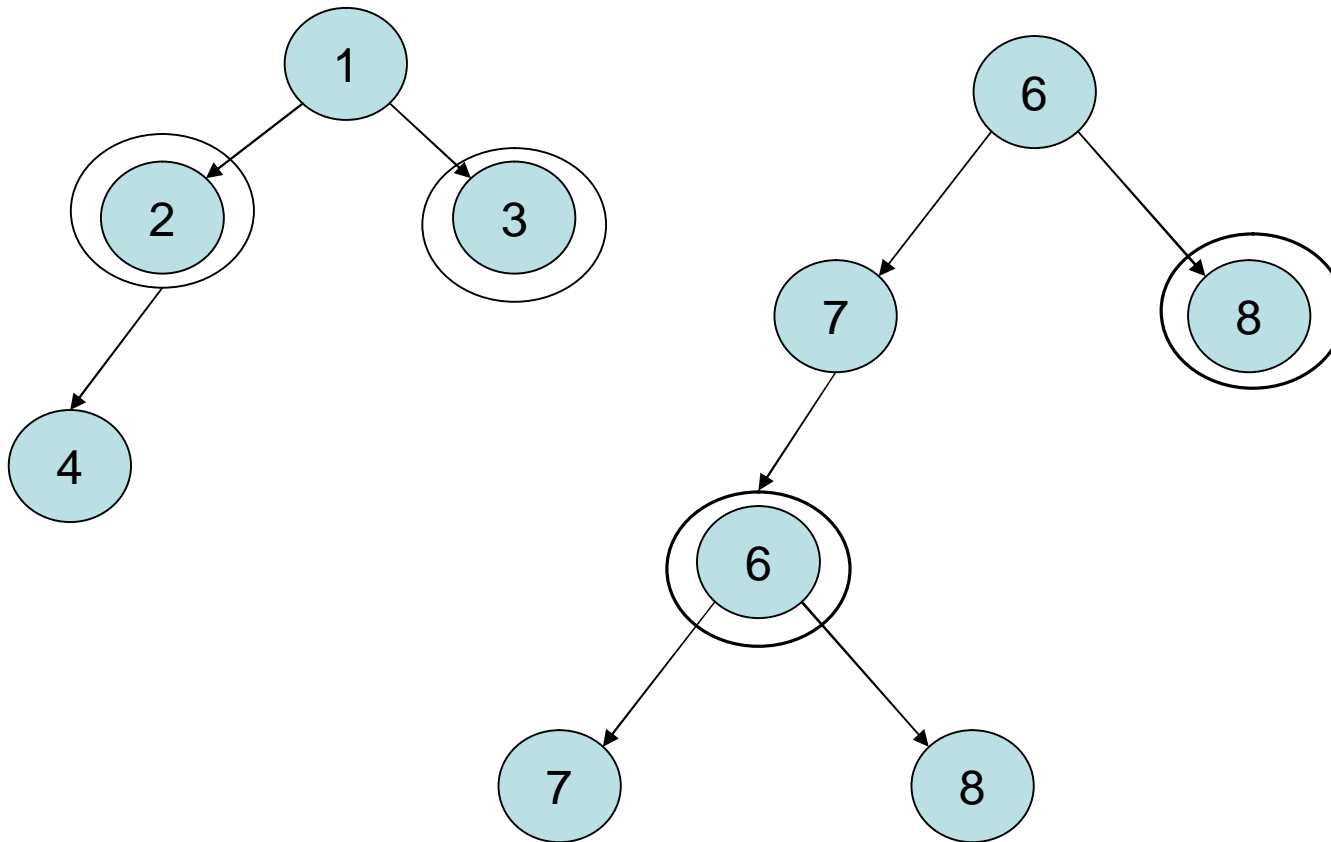
Example geography statistics

- ## 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
  - Countries: CR:1 PE:1 GB:3 FR:1 BO:1 N/A:340 IL:8 VE:1 MY:1 USA:1 TH:1 IN:1 DE:307 EC:1 CA:4 JP:19 CN:1 NL:1 MX:1 BR:1 AR:1 PL:2 CL:1 UY:1 US:34 SG:1 AT:2 ID:1
  - Regions: AMERICAS:61 AP:26 EMEA:471 N/A:164 LATINAMERICA:16

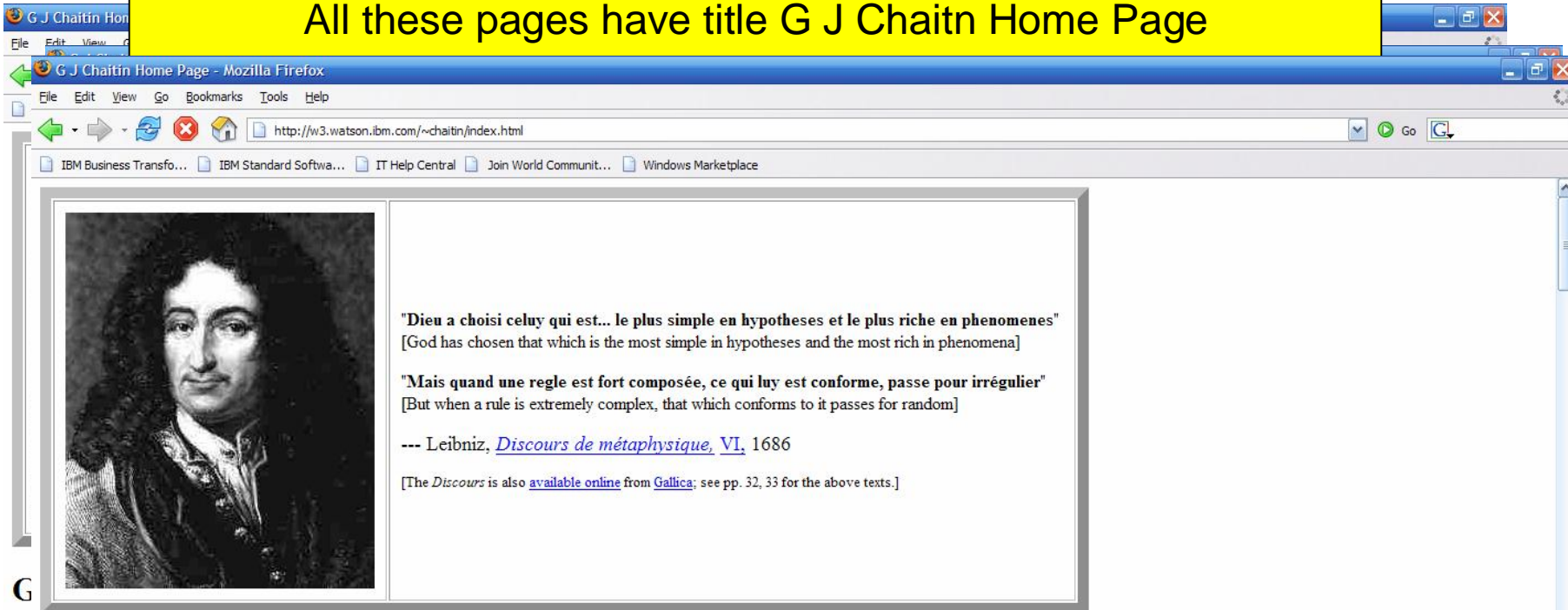# Ranking among types is more important than in-type ranking

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

MRR	all	1kw	2kw	3kw	>3kw
Personal	0.60	0.09	0.93	1	1
A-NamedTitle	0.33	0.38	0	0.30	0
NamedTitle	0.18	0.12	0.16	0.61	1
A-NamedURL	0.27	0.26	0	0.28	0.61
NamedURL	0.22	0.23	0.17	-	-
A-Rooted	0.13	0.17	0	0.03	0
Rooted	0.10	0.05	0.14	0.31	0.61

All these pages have the exact same title IBM Tucson Home

G J Chaitin Home Page - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://w3.watson.ibm.com/~chaitin/index.html

IBM Business Transfo... IBM Standard Softwa... IT Help Central Join World Communit... Windows Marketplace

"Dieu a choisi celuy qui est... le plus simple en hypotheses et le plus riche en phenomenes"
[God has chosen that which is the most simple in hypotheses and the most rich in phenomena]

"Mais quand une regle est fort composée, ce qui luy est conforme, passe pour irrégulier"
[But when a rule is extremely complex, that which conforms to it passes for random]

--- Leibniz, *Discours de métaphysique,* VI, 1686

[The *Discours* is also available online from Gallica; see pp. 32, 33 for the above texts.]

G J Chaitin Home Page

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.

- **E-mail:** chaitin@us.ibm.com

- **Website:** http://www.umcs.maine.edu/~chaitin

- **Mirror:** http://www.cs.auckland.ac.nz/~chaitin

- **Phone:** 914/945-2785

- **Fax:** 914/945-4506

- **Mailing address:**
  IBM Research, P O Box 218
  Yorktown Heights, NY 10598, USA

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**

Transferring data from w3.watson.ibm.com...

Sample navigational panel