

**Thomas Mandl**  
Informationswissenschaft  
Universität Hildesheim  
mandl@uni-hildesheim.de

Fachgruppentreffen  
Fokussierte Suche  
Bamberg 8.5.2008

# Die Klassifikation von geographischen Anfragen in einem Query-Log



# Geographic Information Systems



**Regionaleinheit** Land Bayern - alle Gemeinden

Land  
Reg.-Bezirke  
Regionen  
Kreise  
Gemeinden  
Gemeinde ?

Grenzzlinien  
Reg.-Bezirke  
Regionen  
Kreise

Optionen  
Zoom Aus  
Grenzen

771144 Kühbach  
43 Zuzüge

Gemeinde: Adresse / Homepage

771144 Kühbach

Thema: Fläche, Bevölkerung, Erwerbstätigkeit

Unterthema: Wanderungen

Zuzüge je 1000 Einwohner 2004	43
Fortzüge je 1000 Einwohner 2004	45
Wanderungssaldo insgesamt 2004	-11
Wanderungssaldo 18-25 je 1000 Einwohner 2004	-2
Wanderungssaldo 25-65 je 1000 Einwohner 2004	1
Wanderungssaldo 65 u.m. je 1000 Einwohner 2004	-1

UKRAINA  
SZLOWAKIA  
ALUSZTRA  
ROMANIA  
ORVATORSZAG

Budapest Europa Congress Center, Nemzeti Museum Budapest

Suche  
Route

Karte: statisch | interaktiv  
Sprache / Language: Ändern

Produkte  
News  
Hilfe

stellen

ipest

Center 24H)

IS enburg)

ialmuseum

ipest 8)

stellen

Route finden

Europa

Karte  
Hybrid  
Satellit

Szentendre-Sziget

Budapest

M3/E71

M1/M7/E71

M0/E60/E75

M5

Search term: Unfal

Locality: Frankfurt

Max. results: 250

http://www.wmserver.de:8080 - GeoCLEF - Article details - Mozilla ...

### Der Alptrraum namens Auto

FRANKFURT-NORDWEST (sen). Ein Auto hält neben einem Bürgersteig, der Fahrer läßt den Motor laufen. Der Auspuff spuckt graue Abgase aus und bläst sie direkt auf einen Kinderwagen. Das Kind dreht gequält den Kopf weg. Im dichten Verkehr drängen Autos Fahrräder an den Straßenrand; Autos parken und verbarrikadieren dabei den Bürgersteig; Mütter mit Kinderwagen müssen auf die andere Straßenseite wechseln. Mitarbeiter der Umweltorganisation Greenpeace haben die traurigen Dokumente des deutschen Straßenverkehrs in Hamburg aufgenommen. Doch es hätte überall sein können irgendwo in einer deutschen Stadt. Unsere Städte sind autogerecht, kritisierte Andreas Huber, Greenpeace-Mitarbeiter, der kürzlich auf Einladung der Grünen im Ortsbezirk 9 (Dornbusch, Eschersheim, Ginnheim) im Haus Dornbusch zum Thema Alptrraum Auto? und die Auswirkungen unserer Mobilität referierte. Ein Thema, das die Rahmen der Stadteileile wird...

Fertig

No events

Mandl: Die Klassifikation von geographischen Anfragen in einem Query-Log

- Erweiterung Information Retrieval
  - Anfragen mit geographischen Referenzen
- Cross-Language Evaluation Forum (CLEF)
  - Evaluierungsinitiative für sprachübergreifendes Retrieval
  - <http://www.clef-campaign.org/>
- Track GeoCLEF seit 2004
  - Evaluierung von geographischem Information Retrieval

- Aim: to evaluate retrieval of multilingual documents with an emphasis on geographic search (GIR)
    - *“find me news stories about riots near Dublin”*
- (Fred Gey @ CLEF Workshop 2005)

# Ziele von CLEF

- Infrastruktur für Forschung und Entwicklung cross- und multi-lingualer Information Retrieval Systeme
  - Multilingual Information Retrieval Systeme evaluieren
  - Testsuites entwickeln: Dokumente, Anfragen, Relevanzurteile, stat. Auswertungsverfahren
- CLEF 2008
  - Registration is still possible
  - E.g. Robust Word Sense Disambiguation Task
    - <http://ixa2.si.ehu.es/clirwsd/>



<b>CLEF Year</b>	<b>2005</b>	<b>2006</b>	<b>2007</b>
Nr. of Participants	11	17	13 New: Query Classification Task 6 participants
Nr. of submitted Experiments	117	149	108

# Query Classification Task

- Goal: find geo queries in a log of real queries
- New in 2007
- Organized by Xing Xie (Microsoft Research Asia, Beijing, China)





# Why Query Classification?

- User interface adaptation
  - Result set design should be different for navigational and informational queries (Cutrell & Guan 2007)
  - Geo queries may require map as result (Google)
- Retrieval Performance
  - „More work needs to be done on customizing methods for each topic“ (Harman 2005)
  - Named entities in query could be used for system parameter selection (Mandl & Womser-Hacker 2006)

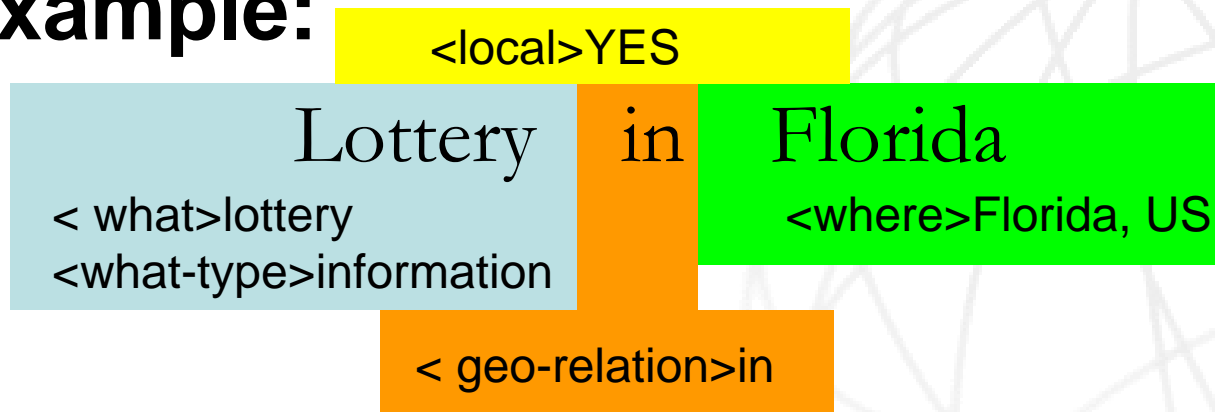


- Query log from the MSN search engine
  - in English
  - 800.000 queries
  - collected August 2006
  - 500 queries were labelled and used for evaluation
    - 100 queries for training
    - 400 for testing

Beaver County times  
BEDDING  
bedroom furniture  
Beer  
Behr Paint  
BEIJING  
BELFAST TELEGRAPH  
belinda  
Belize real estate  
Belkin  
BELL SOUTH  
BELLA VISTA  
BELLAGIO HOTEL  
bellagio las vegas  
BELLE  
belleville  
Belleville News Democrat  
bellevue  
bend it like beckham  
bendigo bank  
Benfica

- Find queries with a geographic scope
  - Extract where component
  - Extract geo-relation-type
  - Extract what component
  - Classify what type {information, yellow page, map}

## Example:



# Geo-relation-types

- 27 classes
- Examples:
  - In
  - On
  - Near
  - Along
  - Distance
  - North\_of
  - North\_west\_of
  - North\_to
  - ...



- 70.000 Queries with Geo Relations
- Davon 50.000 mit Geo Eigennamen

*Li et al. 2007*



Aktuelle Nachrichten Antivir PersonalEditio...

Textmining TM-Taggingtool WS2007/... http://www.br...ing/Classify http://www...line.de/tm/

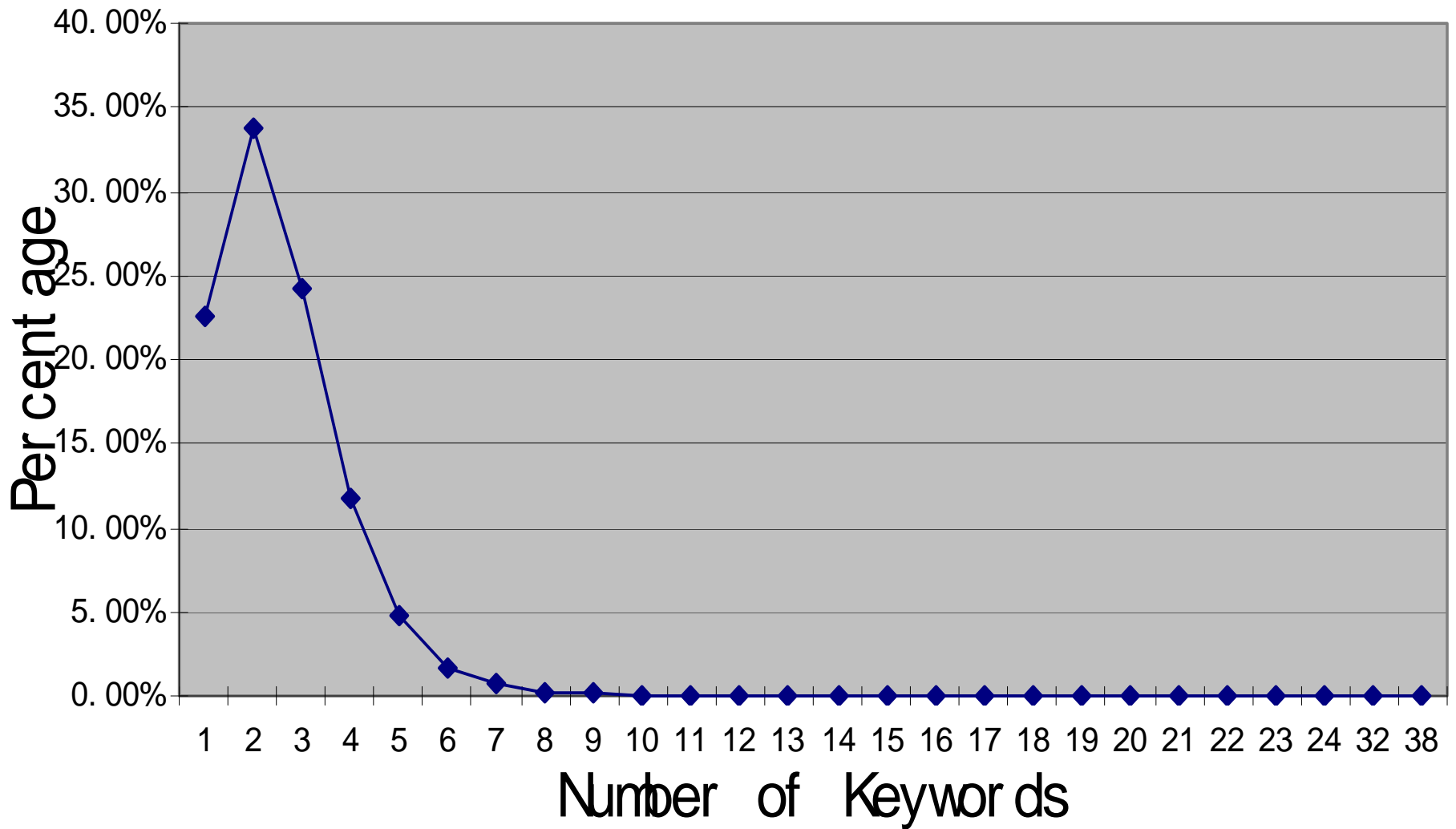
[Keyword-Häufigkeit](#) [Query-Häufigkeit](#) [Keyword-Anzahl](#)

[Download als .csv](#)

Keyword	Häufigkeit
<a href="#">in</a>	48565
<a href="#">of</a>	19437
<a href="#">to</a>	14063
<a href="#">de</a>	13805
<a href="#">new</a>	11241
<a href="#">county</a>	10152
<a href="#">the</a>	9806
<a href="#">hotel</a>	9121
<a href="#">AND</a>	9105
<a href="#">for</a>	8403
<a href="#">estate</a>	8400
<a href="#">real</a>	8135
<a href="#">city</a>	7915
<a href="#">texas</a>	7163
<a href="#">florida</a>	6898
<a href="#">free</a>	6686
<a href="#">home</a>	6554
<a href="#">beach</a>	6453

Angehalten

# Length of Queries



[Keyword-Häufigkeit](#)[Query-Häufigkeit](#)[Keyword-Anzahl](#)[Download als .csv](#)

Warning: Invalid argument supplied for foreach() in  
 /srv/www/vhosts/bressel-online.de/httpdocs/tm/Zend/Cache/Backend/File.php on line 475

Query	Häufigkeit
goOGLE	25
pirates of the Caribbean	19
yAHOO	18
Jesse Mccartney	15
MYSFACE	15
ebAY	13
td Canada trust	13
Yahoo JAPAN	13
naruto	12
mapquest	11
my Space	11
USA Today	11
New york times	11
CATHERINE ZETA JONES	11
Chad michael Murray	11
BANK OF AMERICA	11

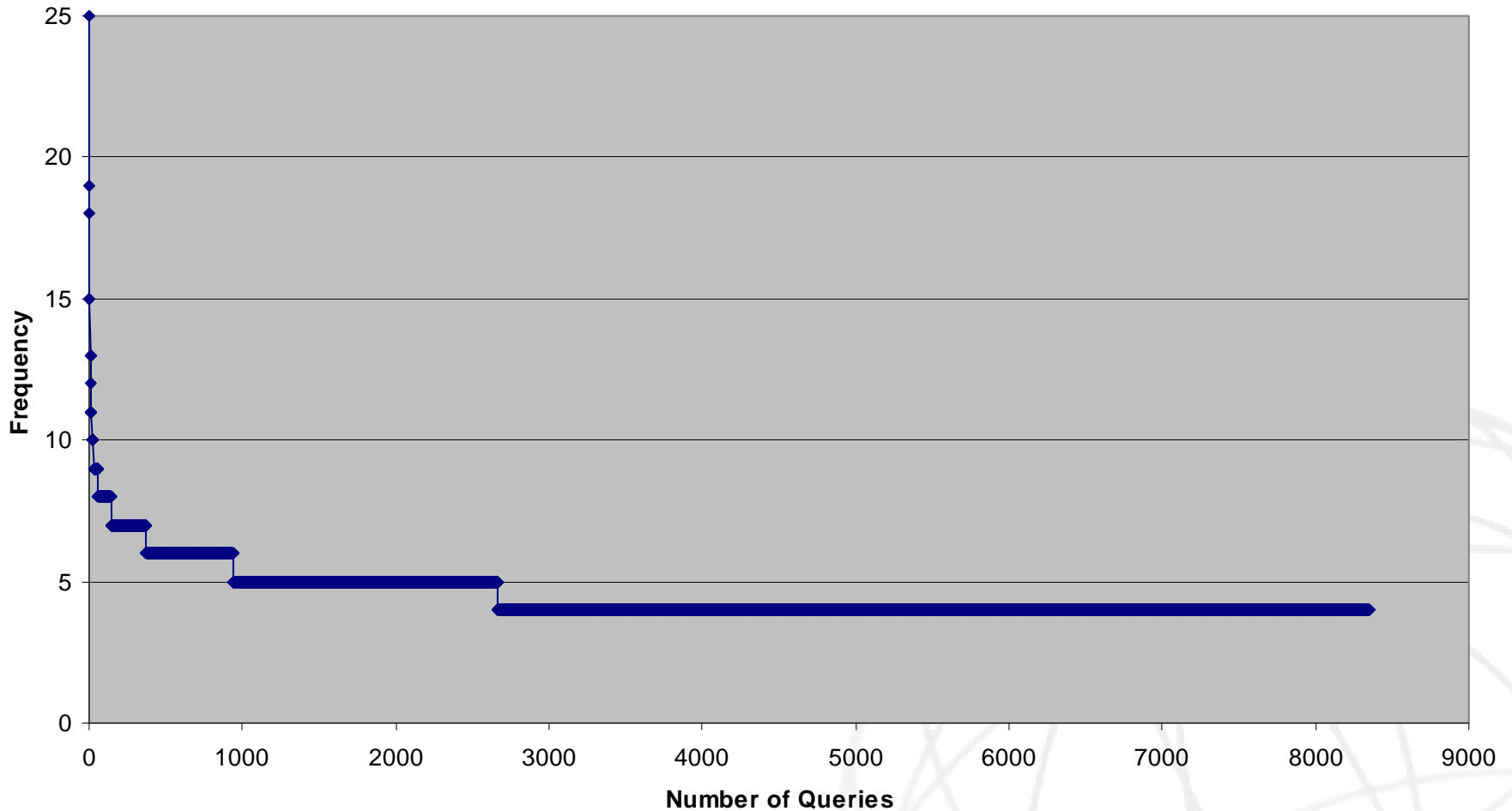

**MSN !**

Many  
 Navigational  
 queries  
 (homepage finding)

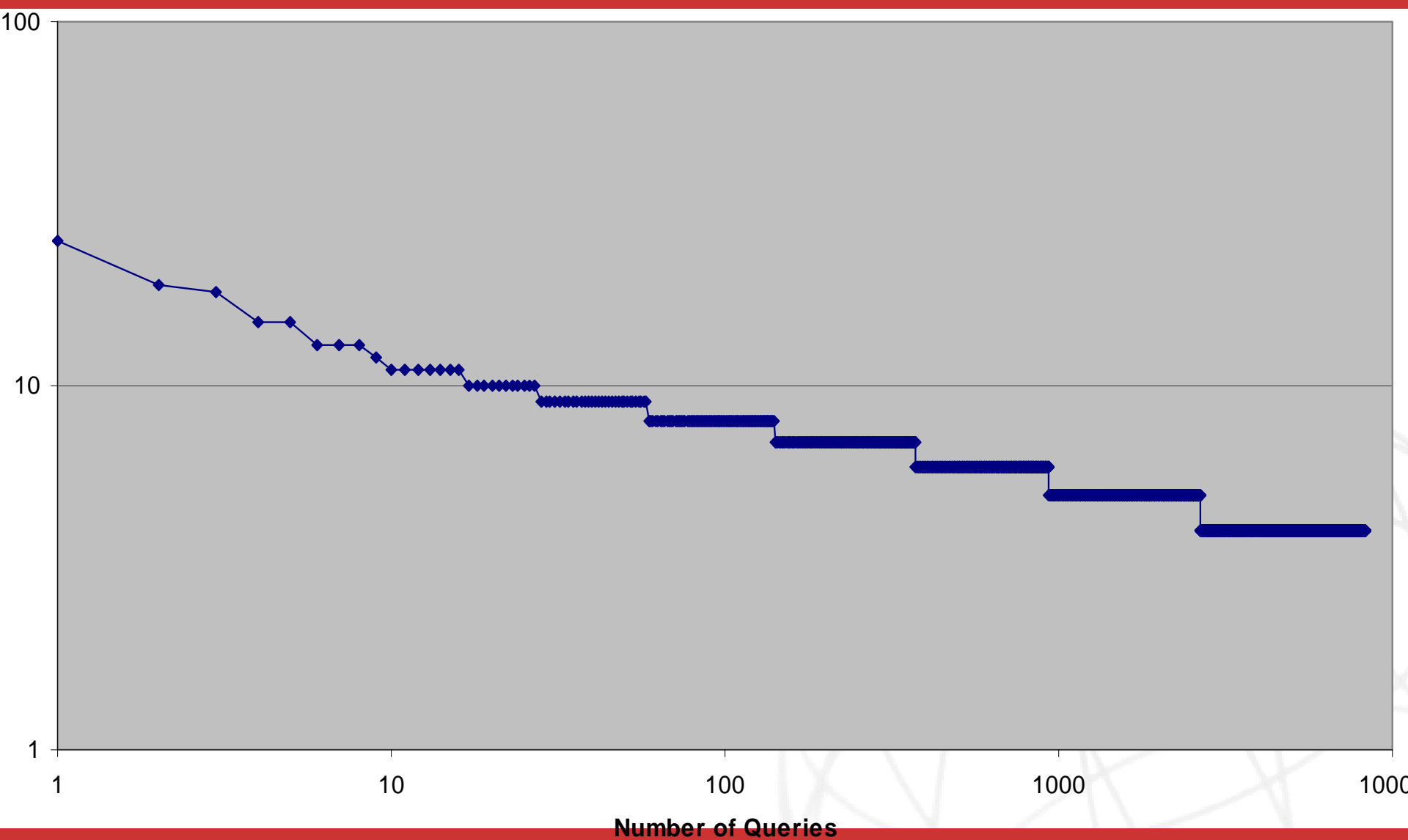
Fertig



# Most frequent Queries



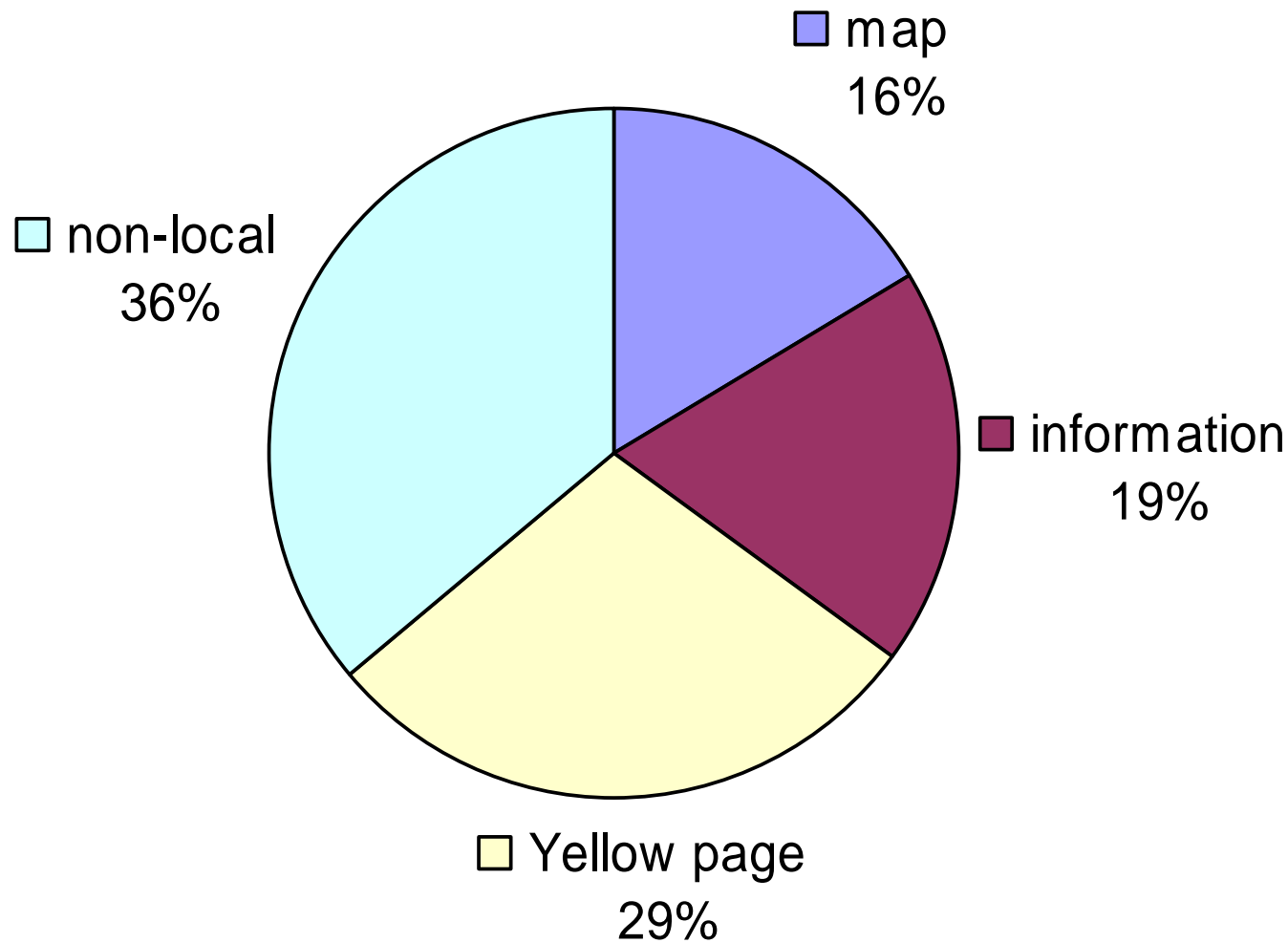
# Most frequent Queries



# Evaluation Set

1. Choose 800 queries randomly from the query set.
2. Remove the typos and the ambiguous queries from the 800 ones manually.
3. Select the queries with special geo-relations from the remainder queries in the query set manually and add them to the evaluation set.
4. Select 500 queries for the final evaluation set.

# Query Types in Evaluation Set



# Comparison to other Studies

- 20% of all Web queries have a geographic context

*(Sanderson & Kohler 2004)*

- Analysis of 1000 Altavista Queries

- Informational 48%
- Navigational 20%
- Transaction oriented 30%

*(Broder 2002)*

# Evaluation Metrics

- Three assessors
  - individually assessed all system answers
  - reached an agreement
- Fully Correct classified query instances
- Recall, precision and combined F1-Score

# Problems

- Ambiguity

- Place names:

*Li et al. 2007*

- Washington

- Place names / other named entities

- Machida (city and actor in Japan)

- Between classes:

*Lana-Serano et al. 2007*

- Atlanta medical (informational or yellow page?)
    - Airport (geo or not?)
    - calabria chat (geo or not?)



Team	Precision	Recall	F1
Ask	<b>0.625</b>	0.258	0.365
Csusm	0.201	0.197	0.199
Linguit	0.112	0.038	0.057
Miracle	0.428	<b>0.566</b>	<b>0.488</b>
Talp	0.222	0.249	0.235
Xldb	0.096	0.08	0.088

- Gazeteers for location identification
  - Large databases of geo names
- Pre-defined Rules
- Issues
  - Low Performance
  - Few training classes for many geo-types

- Small Project in Hildesheim
- Classification
  - Statistical classifier based on keyword frequencies
  - Geo-Name database
  - Tagging Tool to gather more trainings examples

TM-Taggingtool WS2007/08 - Mozilla Firefox

Datei Bearbeiten Ansicht Chronik Lesezeichen Extras Hilfe

http://www.bressel-online.de/tmboris/taggingtool\_beta/ Google

Aktuelle Nachrichten AntiVir PersonalEditio...

Textmining TM-Taggingtool WS2007/08

36) 01NET

37) 01telecharger

38) 02 AND SPRINGDALE AND 26

39) 02 ireland

40) 02 online

41) 02 sensor

42) 02 sim cards

43) 02jam

44) 03t

ireland

kategorisieren als:

Startindex:   
Anzahl der Elemente: 2

05 wisconsin walle

prev. tagging: Für die Kategorisierung wurde kein Text markiert!

Fertig

Textmining - Mozilla Firefox

Datei Bearbeiten Ansicht Chronik Lesezeichen Extras Hilfe

http://www.bressel-online.de/new/index.php

Aktuelle Nachrichten AntiVir PersonalEditio...

Offset	<input type="text" value="0"/>	<input type="button" value="weiter &gt;"/>
Gewichtung	<input type="text" value="12"/> ca. 12	
Anzahl	<input type="text" value="100"/>	
		<input type="button" value="Senden"/>

Location Wahrscheinlichkeit:

- 0 2 5 in review year: **0.062971** [\[geonames\]](#)
- 0 apr: **0.115385** [\[geonames\]](#)
- 0 apr card credit: **0.612360** [\[geonames\]](#)
- 0 apr card credit in uk: **0.612360** [\[geonames\]](#)
- 0 apr card credit low: **0.463732** [\[geonames\]](#)
- 0 apr card credit no: **0.612360** [\[geonames\]](#)
- 0 apr credit: **0.530819** [\[geonames\]](#)
- 0 apr credit card: **0.612360** [\[geonames\]](#)
- 0 Apr Credit Card: **0.050895** [\[geonames\]](#)
- 0 apr credit cards: **0.450195** [\[geonames\]](#)
- 0 apr on balance transfer: **0.090295** [\[geonames\]](#)
- 0 balance card credit transfer: **0.577708** [\[geonames\]](#)
- 0 balance transfer: **0.065217** [\[geonames\]](#)
- 0 balance transfer on credit cards: **0.424010** [\[geonames\]](#)

Fertig

# Negative Example

- „Credit Card“ -> classified as Geo Query
- Positive Training examples
  - „Credit River“
  - „Port Credit“
  - „Card Gulch“
  - „Card Bay“





*Vielen Dank für Ihre  
Aufmerksamkeit*