Analyses des requêtes utilisateurs adaptées à la recherche web

- Journée DAPA - 26 mars 2009 -

Nicolas Stroppa, Romain Vinot

Goal and Outline

Goal

- Final goal:
  - Being able to understand user intent for improving web search
- Intermediate goals:
  - Being able to process the "user queries language"
  - Find suitable (internal) representations for web search queries
  - Develop tools to analyze productions from this language
    - i.e. translate queries into their proper internal representations

Outline of the talk

- Part I - Description of user search queries
- Part II - Presentation of some query analysis tools for web search

Part I
Describing User Queries

User Queries
Examples

20 random queries (from FR)

DEEZER
situation geograhique kosovo
le bon coin
google
the hun's yellow pages
coloriages a imprimer gratuitement
domainenicolas rousset
facebook
www.banque-accord.fr
"sandy koufax"
cheb mami
entretien nautisme la Mézière
"gilles gautier"
www.ca-nord-est.fr
itinéraire Mappy
asse
www.actualite-litteraire.com
fetich finish
skyrock
ph

20 random queries (from UK-Ireland)

cleberson roofing oswestry
car prices
gestalt principles of perception
www.supervalue.ie
pascoa 2009
morgage express
east yorkshire
yahoo
bbc news
heuristic play
hotels nancy
sex.com
farmdata
msn hotmail
02
work function
national rail enquiries
butlins
search genes reunited
facebook

User Queries

A language for web search?

How to formulate queries for web search?

- Boolean
  - (colorier OR coloriage) AND (gratuit)
  - (download AND firefox AND 3.1) AND (NOT 3.0)
- Database-oriented
  - WHERE url='%nikon.com' AND title='%D3X%' AND in_links>0.76 ORDER BY recency
  - WHERE body='%recette%' AND body='%tartiflette%' AND spam<0.05 ORDER BY in_links
- Natural Language
  - Quelle est la date du prochain concert de Radiohead à Bercy ?
  - Je recherche une recette typique de tartiflette
  - Je veux aller sur le site web allocine

User Queries

A language for web search?

How to formulate queries for web search?

- Boolean (dedicated language)
  - (colorier OR coloriage) AND (gratuit)
  - (download AND firefox AND 3.1) AND (NOT 3.0)
- Database-oriented (dedicated language)
  - WHERE url='%nikon.com' AND title='%D3X%' AND in_links>0.76 ORDER BY recency
  - WHERE body='%recette%' AND body='%tartiflette%' AND spam<0.05 ORDER BY in_links
- Natural language (generic language)
  - Quelle est la date du prochain concert de Radiohead à Bercy ?
  - Je recherche une recette typique de tartiflette
  - Je veux aller sur le site web allocine

User Queries

A language for web search?

Dedicated vs. Generic Language

- Using dedicated languages means
  - Educating users
  - ⇔ Modifying the sender
- Using a generic language means
  - Adapting the engine
  - ⇔ Modifying the receiver

A language for web search?

Dedicated vs. Generic Language

- Using a dedicated language means
  - Educating users
  - ⇔ Modifying the sender
- Using a generic language means
  - Adapting the engine
  - ⇔ Modifying the receiver

Remarks

- Trying to explicitly modify users behaviour is not realistic…
- Adapting an engine so as to fully understand natural languages is not any better…
- Reality is actually a mix of both
  - The users and the engine are both constantly making efforts to understand each other! (more on next slides)

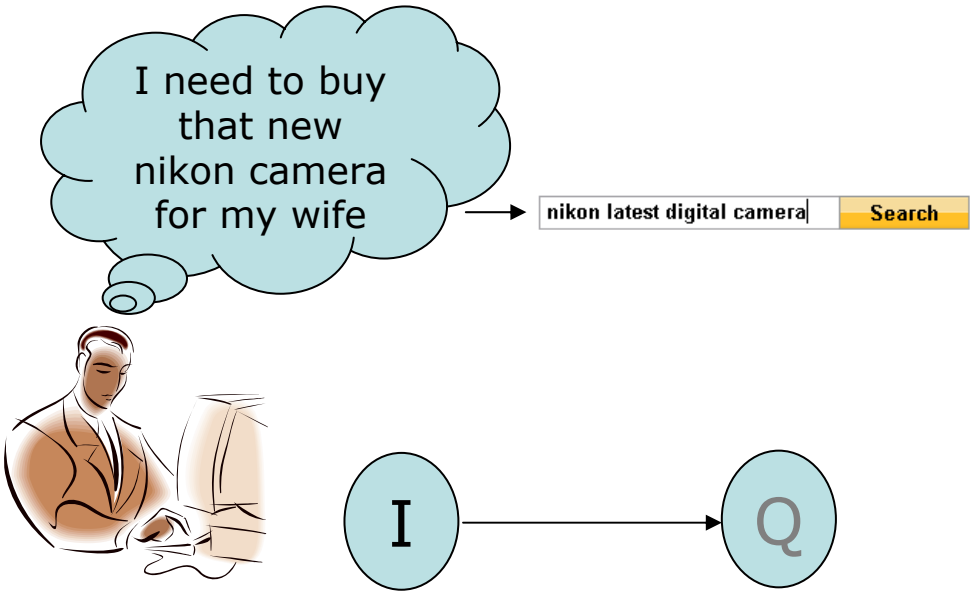
User Queries

User Intent

- I: User intent (hidden variable)

User Queries

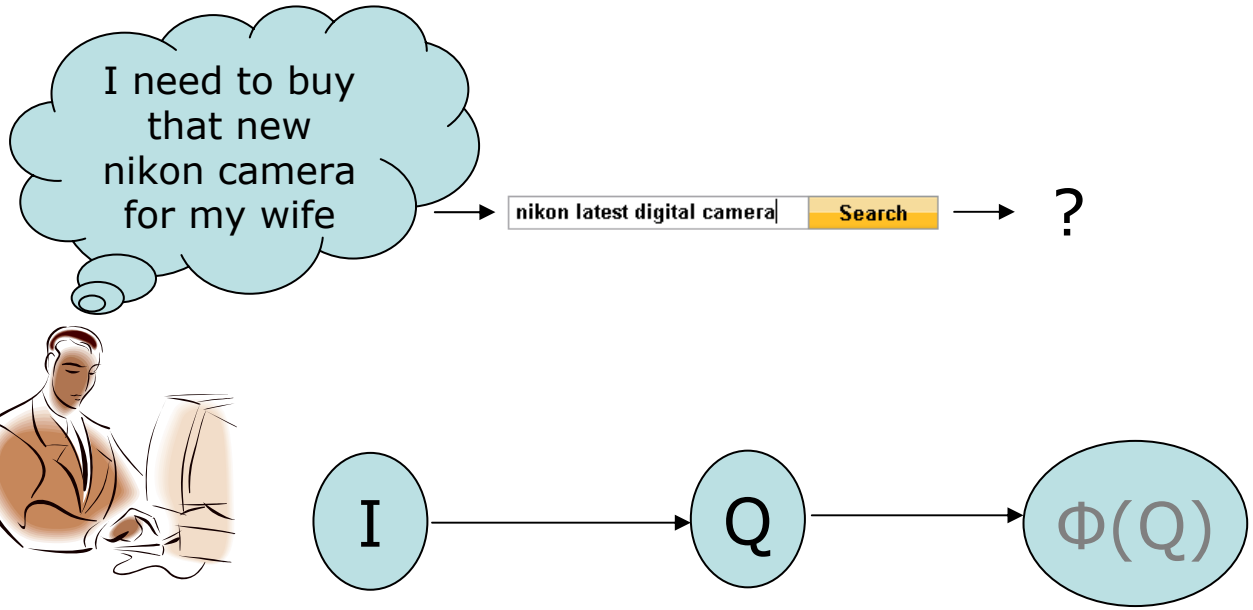
Issued Query



- I: User intent (hidden variable)
- Q: User issued query (observed variable)

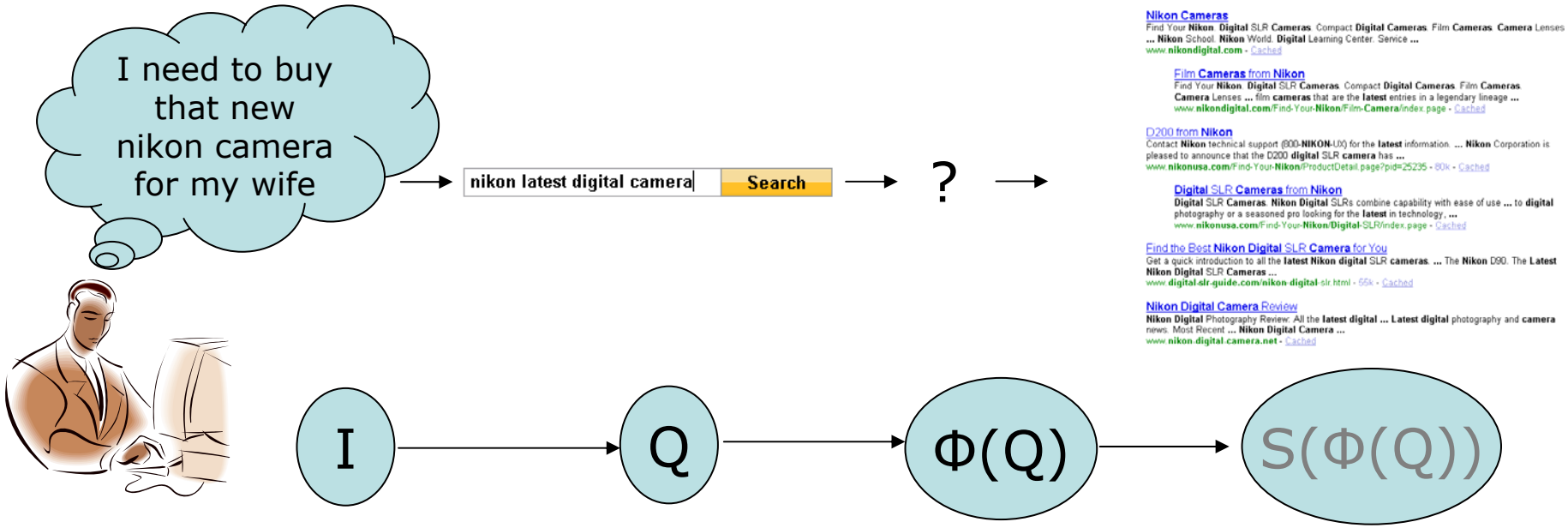
User Queries

Internal Representation



- I: User intent (hidden variable)
- Q: User issued query (observed variable)
- Φ(Q): Internal query representation

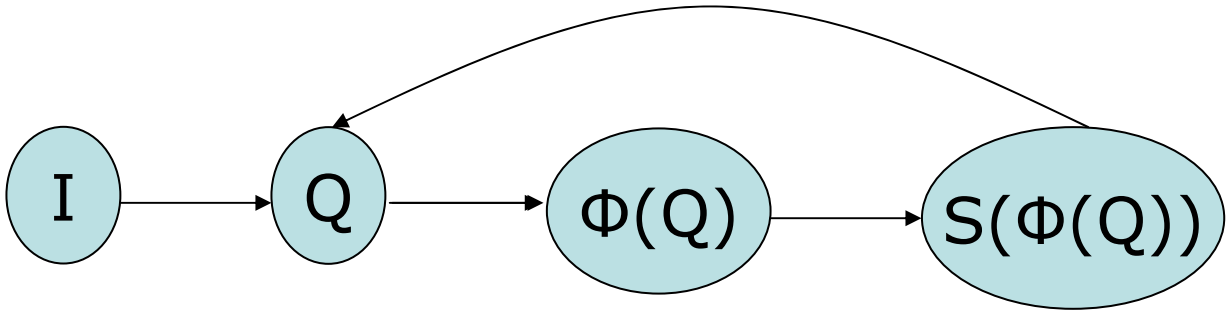
User Queries
Search Results



- I: User intent (hidden variable)
- Q: User issued query (observed variable)
- Φ(Q): Internal query representation
- S(Φ(Q)): Search results

User Queries

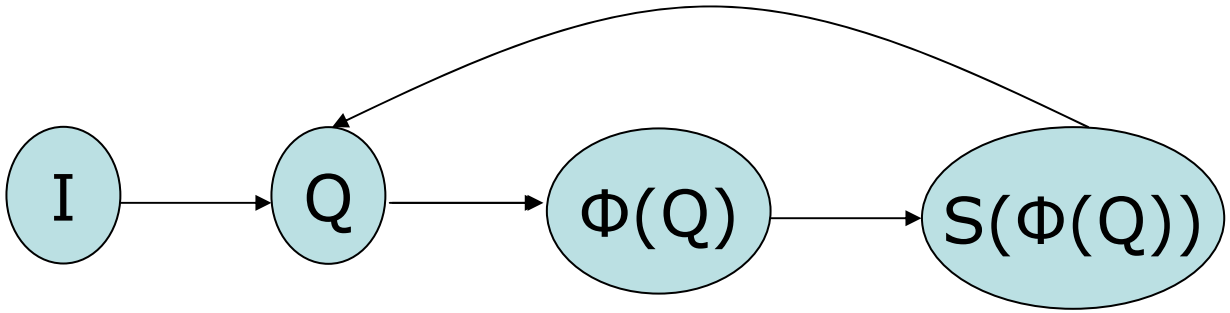
Feedback Loop



- Q depends on I *and* on S:
  - Users tend to learn how to get good results
- Two notable examples:
  - Keyword search:
    - Users have learned to omit stopwords, to limit number of words, etc.
  - Spaces in Japanese:
    - There's no space in written Japanese to delimit "word" boundaries
    - Japanese search users learnt (without any specific "education") that using space was helping the engine tokenize queries => Japanese queries now contain spaces!
- This feedback loop illustrates the user effort to interact with the engine; Φ is the engine effort to understand the user, which is what our work is about (see Part II)!

User Queries

Graphical Model



- Note: all those variables can be modeled as random variables
- I has a time-dependent distribution
  - Users interests are constantly evolving

- Formalizing the objective (R=expected results, S=fixed engine):

$$\mathrm{Min}_{\phi} E_I[\mathrm{Loss}(R(I), S(\phi(\mathrm{Q}(I))))]$$

User Queries

A new language

Some Properties of user queries language

- User/machine interaction created a new language, different from French, English, or SQL
- It shares some properties with natural languages
  - Unrestricted, Zipf law, lots of single occurrences (hapaxes), etc.
- It's different from natural languages:
  - Very light syntax, usually mere sequences of tokens
  - It cannot escape from its interaction with web search
  - Web search use as a bookmark (navigational and domain queries)
  - Lots of variation for same intent (no normative effort)
- Note: a new language implies adapted linguistics tools

User Queries
Examples

20 random queries (from FR)

DEEZER
situation geograhique kosovo
le bon coin
google
the hun's yellow pages
coloriages a imprimer gratuitement
domainenicolas rousset
facebook
www.banque-accord.fr
"sandy koufax"
cheb mami
entretien nautisme la Mézière
"gilles gautier"
www.ca-nord-est.fr
itinéraire Mappy
asse
www.actualite-litteraire.com
fetich finish
skyrock
ph

20 random queries (from UK-Ireland)

cleberson roofing oswestry
car prices
gestalt principles of perception
www.supervalue.ie
pascoa 2009
morgage express
east yorkshire
yahoo
bbc news
heuristic play
hotels nancy
sex.com
farmdata
msn hotmail
02
work function
national rail enquiries
butlins
search genes reunited
facebook

User Queries

Navigational Queries

20 random queries (from FR)

DEEZER
situation geograhique kosovo
le bon coin
google
the hun's yellow pages
coloriages a imprimer gratuitement
domainenicolas rousset
facebook
www.banque-accord.fr
"sandy koufax"
cheb mami
entretien nautisme la Mézière
"gilles gautier"
www.ca-nord-est.fr
itinéraire Mappy
asse
www.actualite-litteraire.com
fetich finish
skyrock
ph

Navigational Queries

20 random queries (from UK-Ireland)

cleberson roofing oswestry
car prices
gestalt principles of perception
www.supervalue.ie
pascoa 2009
morgage express
east yorkshire
yahoo
bbc news
heuristic play
hotels nancy
sex.com
farmdata
msn hotmail
02
work function
national rail enquiries
butlins
search genes reunited
facebook

User Queries

Domain Queries

20 random queries (from FR)

DEEZER
situation geograhique kosovo
le bon coin
google
the hun's yellow pages
coloriages a imprimer gratuitement
domainenicolas rousset
facebook

www.banque-accord.fr

"sandy koufax"
cheb mami
entretien nautisme la Mézi
"gilles gautier"

www.ca-nord-est.fr

itinéraire Mappy
asse

www.actualite-litteraire.com

fetich finish
skyrock
ph

Domain Queries

20 random queries (from UK-Ireland)

cleberson roofing oswestry
car prices
gestalt principles of perception

www.supervalue.ie

pascoa 2009
morgage express
east yorkshire
yahoo
bbc news
heuristic play
hotels nancy

sex.com

farmdata
msn hotmail
02
work function
national rail enquiries
butlins
search genes reunited
facebook

Part II
Analyzing Queries

Analyzing Queries

Objective

- Build a suitable (internal) representation of user queries for web search (Φ)
  - => Engine effort to understand the user intent

Expected property

- Identical intents should lead to the same results
  - One way to ensure that:
    - same intent => same internal representation

Taken approach

- A user intent is defined as an "equivalence class" on queries
  - All possible formulations of the same intent belong to the same class
  - (=> A query should be safely replaced with any variant formulation)
- Advantage
  - Does not need any explicit representation for intent
- Limitation
  - Due to ambiguity, we cannot properly define equivalence classes

User Queries

Internal Representation

Our Model

- Φ(q) is the (weighted) list of queries that can be issued given the intent behind q
- We call this list the "variants" of q
- This list is the input to the search engine
  - The assumption is that we have a search engine that can process these "raw" queries, and does not do so bad a job at it
  - The model is built on top of an existing search engine, and thus cannot deeply "break" it

User Queries

Variants

Variants considered

- Due to ambiguity issue, we prefer to keep a restricted view on the notion of variants
  - Smaller equivalence classes ⇔ Focus on precision
  - Useful approximation: focus only on "promising" variants, i.e. those that are likely to bring relevant results (e.g. it may be legitimate to discard variants ranked lower in the list)
- Examples of types of variants considered:
  - Orthographic variants
  - Morphological variants
  - *Semantic variants (almost not covered in this talk)*
  - *Pragmatic variants (not covered in this talk)*

Analyzing Queries

Orthographic variants

Use case

- Navigational query
- Target site: www.lequipe.fr

Examples of issued queries

;'equipe
equipe
equipe fr
equipe.fr
l equipe
l' equipe
l'2quipe
l'equipe
l'equipe.
l'equipe.fr
lequipe
lequipe.fr
lequipelequipe
lequipes

Analyzing Queries

Orthographic variants

Use case

- Navigational query
- Target site: www.youtube.fr (or www.youtube.com)

Examples of issued queries

utub	yputube	you tube*
utube	yu tube	you tubee
yotub	fyoutube	you tuber
utubes	tou tube	you tubes
u tube	wyoutube	your tube
y tube	you tub	youtubess
yootub	you tube	yuo tubes
youtub	you tubr	+you +tube
yuotub	youttube	www yutube
yutube	youtubx	you +tubes
yu tube	youtub e	youtttube
toutube	youtubee	youtub.com
yo tube	youtubes	youtube fr
yoo tub	youtubze	youtube.fr
yootube	yoy tube	youtubrese
yooutub	yuo tube	www.youtube
yotubes	you tube0	youtube.com
you tub	you ntube	yuotube.com
youtube		

Orthographic variants

Misspellings / Orthographic Variants

- Goal:
  - Find "useful" orthographic variants
  - The notion of usefulness is defined only with respect to the quality of returned documents (i.e. user satisfaction)
- Concerned queries
  - ~10% of user queries
    - Feedback loop: users have learned that misspelled queries were automatically corrected…
  - Covered cases: typos, misspellings, spaces (split/join), apostrophes, punctuations
- Often an explicit messaging

Orthographic variants

Model

- Supervised machine-learned model that makes use of
  - Query/Suggestion features
    - Occurrences, reformulations in logs, edit-distance, etc.
  - Click features
  - Search result features
    - Number of hits, etc.
- Trained and tested on human annotated data
- Precision-oriented: a miss (false negative) is better than a mistake (false positive)

Orthographic variants

Not so easy cases…

- xhamster => hamster
- jukebo => jukebox
- starbooker => starcooker
- pmum => pmu
- televysion => television
- tecktonique => tectonique
- metzanine => mezzanine
- etc.

Orthographic variants

Not so easy cases…

- xhamster => hamster (left=popular adult site)
- jukebo => jukebox (left=popular clip site)
- starbooker => starcooker (left=social site, right=restaurant)
- pmum => pmu (left=pari mutuel urbain maroc)
- televysion => television (left=online tv)
- tecktonique => tectonique (left=danse, right=science)
- metzanine => mezzanine (left=zone d'activite commerciale de Metz)

- and very little context to disambiguate…

Analyzing Queries

Orthographic variants

Not so easy cases…

- **xhamster => hamster** (left=popular adult site)
- **jukebo => jukebox** (left=popular clip site)
- starbooker => starcooker (left=social site, right=restaurant)
- pmum => pmu (left=pari mutuel urbain maroc)
- **televysion => television** (left=online tv)
- tecktonique => tectonique (left=danse, right=science)
- **metzanine => mezzanine** (left=zone d'activite commerciale de Metz)

site name or proper name
close to a common noun

- Note: very high number of "entities" in user queries (sites, person names, locations, product names, etc.)

Orthographic variants

Not so easy cases…

- xhamster => hamster (left=popular adult site)
- jukebo => jukebox (left=popular clip site)
- starbooker => starcooker (left=social site, right=restaurant)
- **pmum => pmu** (left=pari mutuel urbain maroc)
- televysion => television (left=online tv)
- tecktonique => tectonique (left=danse, right=science)
- metzanine => mezzanine (left=zone d'activite commerciale de Metz)

close abbreviations

Morphological Variants

Examples of Morphological Variants

- French:
  - hotel club du soleil ⇔ hotel(s) club(s) du soleil
  - recettes cuisine ⇔ recette(s) cuisine(s)
  - film gratuit a regarder sur PC ⇔ film(s) gratuit(s) a regarder sur PC
  - centre de recherche européen + Italie ⇔ centre(s) de recherche(s) européen(s) + Italie
  - salle de bains ⇔ salle(s) de bain(s)
  - agence immobilière ⇔ agence(s) immobilière(s)
  - Ajouter feminin/masculin
- German:
  - kleider in kataloge => kleide(r) in katalog(e)
  - auto lackieren preis => auto lackieren preis(e)
  - schlüssel langen => schlüssel langen/lang/langer
- English:
  - english bulldogs => english bulldog(s)
  - sorting mail on the train => sort(ing) mail on the train

Analyzing Queries

Morphological variants

Morphological Variants

- Goal:
  - Find "useful" morphological variants
  - The notion of usefulness is defined only with respect to the quality of returned documents (i.e. user satisfaction)
- Concerned queries
  - Covered cases: mostly nouns and adjective inflection (very few verbs)
- No explicit messaging: (query=salle de bains)

Salle de Bains : le 1er site de la **salle de bain** pour votre instalation ...
100 aménagements **de salle de bains**, des idées de grands créateurs, des **salles** d'exposition près de chez vous, solution, conseils et adresses utiles,
www.**salledebains.fr** - En cache

CUISINELLA : Découvrez nos **salles de bains**
Découvrez Cuisinella, ses cuisines et ses **salles de bains**, ses avantages et construisez en ligne, accompagné d'une ou d'un assistant, votre liste d'envie. Simple, ...
salle-de-bain.cuisinella.com/salle-de-bain.php - En cache

Salle de bain : Douche, Baignoire et meubles **de salle de bain** THALASSOR
salle de bain. Découvrez notre gamme **de salle de bain**.
www.**salle-bain.com** - En cache

Morphological variants

Model

- Unsupervised models
  - List of language-specific morphological variants
  - Language models built with query logs (i.e. model the ‘user query language’)
- Mostly unsupervised but some internal parameters are tuned for optimizing relevance metrics (e.g. DCG), which require some human judgments

Morphological variants

Not so easy cases…

- Relatively easy cases:
  - singular/plural variations for nouns
- But:
  - les 7 mercenaires
  - le renard et la cigogne
  - la banque populaire
  - la mauvaise réputation
  - etc.
- The high number of entities in user queries makes it harder to determine if two queries are variants of the same intent

Analyzing Queries

Semantic Variants

Semantic variants

- query=hisd schools

HISD Connect - Home
HISD, HCC Break Ground on New East Early College High **School** ... List of **Schools**. Directions to Facilities. **HISD** RSS Feeds. Houston Real Men Read ...
www.**hisd.org** - 102k - Cached

School Calendar
Region Information
Hisdconnect
Hisd News Today (TV)
Doing Business With **Hisd**
News
Board

more results from hisd.org »

HISD Connect - **Schools**
HISD, HCC Break Ground on New East Early College High **School** ... Request for Proposals: **HISD** Superintendent Search. Volunteers in Public **Schools** (VIPS) ...
houstonisd.org/HISD ConnectDS/windex.jsp?vgnextoid=924c2f796138c010... - Cached

Houston Independent School District - Wikipedia, the free encyclopedia
History | Bilingual... | Student body | Governance

The Houston Independent **School** District (**HISD**) is the largest public **school** system in Texas and the seventh-largest in the United States. Houston ISD serves as a community **school** district for most of the city of Houston and...
en.wikipedia.org/wiki/**Houston_Independent_School_District** - 115k - Cached

- query=US Marine

United States Marine Corps
Official U.S. **Marine** Corps recruitment site features a history about the **Marines** and everything you need to know before joining the **Marines**.
www.**marines.com** - Cached

Semper Fidelis
Silent Drill
Reserves
Usmc Birthday
Sniper
Dress Blues
Bases

more results from marines.com »

United States Marine Corps (USMC)
Official **Marine** Corps web site. Includes information on recruiting, units, careers, and more.
www.**usmc.mil** - 68k - Cached

Headquarters **Marine** Corps
Marine Corps Orders and Directives. Pubs Library Management System ... **The United States Marine** Corps (USMC) is a branch of the U.S. military responsible for ...
www.**usmc.mil**/units/hqmc/Pages/default.aspx - 83k - Cached

Conclusion

Conclusion

- Processing a query has no use per se
- What's important is the intent behind the query
  - User expectations: from "give me what I said" to "give me what I want" (Amit Singhal, Google)
- We need to build tools for understanding this intent, which is accessible only indirectly through a 'user query language'
- Identifying query variation given an intent is a component built with those considerations in mind, and bring significant improvements in user experience and relevance gain


LIFE ENGINE™