User-aware page classification in a search engine

Rachel Aires
Diana Santos
Sandra Aluísio

Problem setting: shortcomings of Web IR

- The problem is no longer just to find results,
  - but also to navigate among their massive number in a meaningful way
- Web users are extremely varied and use the Web for all kinds of purposes and activities,
  - so standard user testing in a usability lab with a sample of the population is obviously out of the question

Problem setting: shortcomings of Web IR

- These two factors have led researchers to
  - investigate other properties of Web documents other then their content (such as style, language quality, text genre, authority, novelty, maintenance/update rate)
  - offer personalization solutions for specific kinds of search, for specific kinds of users, for specific kinds of content, as well as try to study the users and their behaviour using unobtrusive techniques
- Our work follows these lines...

Use of style to classify search results according to user needs

- In two ways:
  - Classifying texts according to a taxonomy of seven general user needs
  - Classifying texts in binary personalised taxonomies related to text types
- We were inspired by Karlgren's (2000) studies,
  - and this work can be seen as a validation and confirmation of his original hypotheses, for Portuguese
- Other people have investigated automatic genre classification for other languages,
  - as Stamatatos et al. (2000) for Modern Greek
- We were also inspired by Biber's (1988) empirical investigation of co-occurrence patterns among linguistic features and,
  - in fact, most of our features were directly inspired by (adapted from) those used by Biber

The main hypotheses

- It is possible to distinguish among different users' needs
- In general, different pages satisfy different users' needs
- It is possible in a majority of cases to evaluate whether a particular page or site was designed to satisfy (or not) a given user's need
- Users are aware of their needs (if asked to distinguish among several)
- It is possible to automatically classify pages according to user's needs using features similar to those used in other works to classify texts according to genres
- Users have higher satisfaction when getting results' pages clustered around user's needs

Technical/philosophical User-related

A taxonomy of user needs

When searching the Web, a user may be trying to find:

- Definitions, or learning how or why something happens;
- Procedures: how to do something;
- Comprehensive presentations or surveys about a given topic;
- News on a specific subject;
- Information about a specific person, company or organization;
- A specific webpage which has been visited before;
- The address (URL) of specific online services.
- This taxonomy was inspired in Broder (2002) and in a detailed analysis of TodoBr logs

The first three technical hypotheses

- The first three have received some confirmation by the compilation of the *Yes, user!* corpus of 1,703 Brazilian Web pages
  - a typology of seven user's needs was devised after careful qualitative consideration of the logs of a real Web search engine (Todobr)
  - 4 different people looked at Web pages and classified them according to a set of precise guidelines with examples;
  - although there was considerable overlap (many pages satisfied more than one need), no pages satisfied at once every need
- See paper at CL2005 and www.linguateca.pt/Repositorio/yesuser.html

Classifying pages according to user needs: confirming the last technical hypothesis

- In 6 user needs (removing the "known page item") plus "others", accuracy was 79.38% using 108 features and Logistic Model Trees (LMT)
- Distinguishing solely between information and service, the accuracy was 91.19%
- For comparison, we used a sub-corpus of Lácio-Ref Project and its genre and type of text classification:
  - In automatic genre classification (scientific, informative, law, literary, and instructional), accuracy was 97.21% using 51 style markers and LMT, or 94.87% while considering informative texts extracted from the Web
  - In automatic text types classification (29 types), accuracy was 83.95%
- More details in our paper

But what about the user related hypotheses?

- 1. Are users aware of their needs (if asked to distinguish among several)?
- 2. Do users get more satisfied when the results' pages are presented to them clustered around user's needs?
- 3. Are users able to create meaningful binary corpora for their own specialized needs?

- How can we test these hypotheses?

1. Are users aware of their needs if asked to distinguish among several?

- A questionnaire was delivered to prospective users:
  - undergraduate students of Computer Science, Linguistics, Medicine, and to graduate students of a Photography course
- The users were asked:
  - about the usefulness of our taxonomy for their search on the web and about the usefulness of genre based classification
  - whether they would be interested on a type of personalized classification even though they would have to create a corpus
  - which type of user needs they would choose to complete seven tasks (to confirm whether they hade really understood the palette)

Are users aware of their needs if asked to distinguish among several?

- Of the 63 students, 61 considered the user needs classification useful
- All 41 people who reported to have frequent problems on their searches said they were interested on the personalized search option
- 14 students made mistakes while choosing a type of user need to complete one of the following tasks, namely to find:
  - The Palmeiras homepage
  - A review of a Skank CD or concert
  - A list of hotels in Araraquara
  - The recipe of a cake
  - A website to send electronic cards
  - The causes of the greenhouse effect
  - How to contact the headquarters of Embrapa in São Carlos

Leva-e-traz: desktop meta searcher prototype

2. Are users more satisfied if their search results are presented by user needs?

- 10 users were asked to read a help page about the information needs covered, and use the desktop meta searcher Leva-e-Traz
- They were requested to perform 6 tasks:
  - Buy the book or movie *Wuthering Heights*
  - The meaning of *bode* (goat) and *carneiro* (sheep) in parapsychology
  - Causes for domestic fires
  - Methods of prevention of rabies in human beings
  - Information about misery in Africa
  - What is and how to do *mensalão*
- The need was wrongly chosen only once

2. Are users more satisfied if their search result is presented by user need?

- 6 tasks: 3 of them without classification and 3 with results classified by user needs
- For each task we looked at:
  - the number of times the first result was relevant (29/18),
  - the number of times the first result the user think was relevant really was (20/13) and
  - how many of the results the user judged as relevant were relevant
- 7 users said that the classification was useful
- In 10 cases users said that they found results they would not find without the classification (33%)
- In 6 cases the users said the classification leaded them to judgment errors (20%)

3. Are users able to use the personalized option?

- Do users understand what is the personalized option?
  - Do they understand that we do not deal with subject problems?
- Are they able to create training corpus?
- Can we classify texts reliably in personalized needs?
- Will users still be interested in this option after trying it?

The only way to try this is actually asking users to do that...

Yes, user! binary corpora

- **We have initially created 3 domain specific binary corpora:**
  - Law texts for experts versus texts for laymen in Portuguese
  - Law domain texts for experts versus texts for laymen in English
  - Whether E-commerce pages described products on sale or not
- **6 users have created 7 specific corpora. They were interested in:**
  - Theoretical texts about philosophy of language and (2) Portuguese linguistics
  - Facts about Fado
  - Descriptions of historical themes
  - Glossaries, recipes and techniques Japanese cookery
  - History and facts about Surrealism
  - Technical texts about aeroplanes and aviation

Classifying pages according to binary problems

- Law domain texts for experts versus texts for laymen in Portuguese
  - Accuracy: 83.2% using 46 style markers and LMT
- Law domain texts for experts versus texts for laymen in English
  - Accuracy: 94% using 52 features and Sequential Minimal Optimization (SMO)
- Finding whether E-commerce pages described products on sale or not
  - Accuracy: 89.91% using 46 style markers and LMT

Classifying pages according to users' binary problems

- Around 6 hours to create each corpus
- Accuracy using 46 style markers and SMO:
  - P1: 87.38%
  - P2: 85%
  - P3: 79.9%
  - P4: 85.37%
  - P5: 87.5%
  - P6: 65.5%
  - P7: 73%
- The users who created the corpora said they would do it again for their specific needs

Concluding remarks

- It is feasible to classify Web pages according to the kinds of goals users have, as well as according to genre or type of text dimensions
- (Some) users are willing to create training corpora to facilitate search about their specific interests
- Users seem to be positive to “clustering” their results according to a user needs classification
- There is a lot to be done in order to include this type of classification in working systems
- Most experiments will be flawed one way or the other, but it is better than having no user tests at all! At least they are inspiring...
- It is time to people do research on their own languages and Webs – it is not necessarily the case that everything that applies to English applies to all the other languages