Business plan for the Browsing and Matching Project

Marc Wilhelm Küster, Main Editor

What this project team sets out to achieve

Introduction

The European Commission decided last year to found a project team that is to investigate the scope of further research on *European matching rules*. It will work under the supervision of CEN TC304 and to which it will deliver a draft study report by April 1999.

The problem of string matching − that is the question if two strings are to be considered equal − can be arbitrarily complex. Simple binary comparison between two strings is, of course, well understood and efficient, but fails to bridge even the most basic divergencies between semantically identical expressions,[1] the most simple of which might be that the data is stored in two different code pages.

The task soon becomes more ambitious. Human readers[2] will naturally recognize that *sing, sang, song*[3] are just three tenses of the very same verb, just as *œil* and *yeux* differ only with respect to number. They will also not mix the German word *Boot* with its English homograph of completely different meaning,[4] whereas they understand at once that *Pericles*, *Perikles* and *Περικλῆς* are really one and the same person[5] and that *browsing* and *scanning* can be synonyms[6] in some contexts but not in others.[7]

For English with its fairly limited number of irregular verbs and its otherwise regular construction of derived forms some of these problems can still be dealt with relatively easily in comparison with most other European languages where word formation is far more complex.

Intelligent matching becomes more relevant by the day in an increasingly interconnected world in which access to information is of primordial importance. However, even now cutting edge, high-performance search engines are not able to fulfil but the most basic of demands of European citizens. Modern databases try to defuse this situation by applying a rigid system of categorization via the introduction of some higher level protocol,[8] but this leaves aside the vast amount of existing information.

Scope of this project

This study aims »to investigate the European needs and problems with searching and browsing in relation to character sets, transliteration matching and ordering rules and other cultural specific elements.«[9] It will present an overview about current research projects, their approaches and results, try to isolate major problems and suggest some tentative paths towards their solution. It will try to establish where solutions seem within easy reach, were approaches are likely to be fruitful in medium term and where no solutions can be expected in the foreseeable future.

The project team will collaborate closely not only with research institutes but also with major enterprises[10] working in the field. Search engines require here special attention. Furthermore, it will get in contact with relevant consortia such as the W3C and FIPA.

---

[1] »Expression« shall be used as an equivalent to »string of letters« on which this study shall concentrate
[2] assuming that they are literate in the language(s) in question
[3] problem of irregular verb and nouns forms. Declination and conjugation come in here
[4] problem of disambiguation
[5] problems of non standardized transliteration and of the handling of different scripts. Resolution of spelling ambiguities (e. g. *Göthe* vs. *Goethe*)
[6] putting to use of thesauri
[7] question of matching on natural languages
[8] cf. in this context the diverse digital library projects
[9] CEN/TC304 N739 (19.9.1997), P27,1
[10] this would encompass major search engines such as AltaVista and Yahoo and major database vendors such as Oracle and Siemens

The aim of this study is to give an overview of current practice in Europe, of ongoing research projects and of desiderata.

In accordance with the schedule the project team shall deliver a draft report by week 16 in 1999 to be presented to the plenary of CEN TC304. This report can purposefully only serve as a starting point for further research which, subject to approval, this project team aims to pursue thereafter.

Group

For the task of preparing this report CEN/TC304 established a project team of three experts in the field. Everybody is very welcome to contact the members of the project team.

Advising editor

John Clews
Director of SESAME Computer Projects
8 Avenue Road
Harrogate
HG2 7PG
United Kingdom
Tel: +44–1423 888 432
Email: matching@sesame.demon.co.uk

Advising editor

Hans van der Laan
NNI
W. F. Hermanszijde 3
NL–2353–LT Leiderdorp
The Netherlands
Tel.: +31 71 541 6431
Email: vdlaan@pobox.leidenuniv.nl

Main editor

Marc Wilhelm Küster
Computing centre of the University of Tübingen
Dept. Literary and Documentary Data Processing
Wächterstraße 76
D–72074 Tübingen
Germany
Tel.: +49–7071–29 70348; +49–7071–29 70201
Fax: +49–7071–29 5912
Email: kuester@zdv.uni-tuebingen.de

Working methods and schedule

Division of work

− Establishing of a dedicated E-mail list and maintaining of the project's Web site: main editor

- Continuous discussion of results: all pt members
- Contact with relevant institutions: all pt members
- Writing of the report: main editor
- Reviewing of the report: all pt members

Schedule

Due to the unforeseeable changes in the setup of the project team the original schedule is no longer tenable. The project team hopes to be able to adhere to the following time frame.

- Contact with relevant institutions: From the start
- Continuous discussion among team members: From the start
- Open discussion within TC304: From March 1998 onwards
- Presentation of first draft for the Tübingen plenary of CEN/TC304 (week 16 in 1999)
- Presentation of the final draft for the following plenary of CEN/TC304 (week XX in 1999)