PÁGICO: Evaluating Wikipedia-based information retrieval in Portuguese
SIGA: a System to Manage Information Retrieval Evaluations

LUÍS COSTA, CLÁUDIA FREITAS, CRISTINA MOTA, DIANA SANTOS AND ALBERTO SIMÕES

http://www.linguateca.pt

MOTIVATION AND TASK

Is it possible to develop better systems to answer realistic user needs, searching for answers to a particular topic in Wikipedia? Is Wikipedia in Portuguese good enough to provide information on lusophone topics? Can we learn from watching people trying to answer them? Is competition or cooperation between human and automatic participants worth indulging in?

PT WIKIPEDIA IN 150 TOPICS

Information needs related to Portuguese-speaking countries and their history, with enough coverage in Wikipedia, and not easily browsable through simple categories or infoboxes, spanning areas from History (50) through Geography (26) and Music (19) to Mathematics (1) and Geology (2). How to assess the answers and their justifications was often quite difficult.

PÁGICO COLLECTION

- based on the 25 April 2011 wikipedia snapshot;
- converted to XHTML using:
  - `mwlib` for the markup conversion;
  - `MediaWiki::DumpFile` to control the snapshot parsing;
  - in-house tools to manage macro expansion;

Collection constitution:

Page type	Total docs
Template pages	32 900
Disambiguation pages	5 006
Redirection pages	574 077
Multimedia pages	9 678
Article pages	856 005

EVAL MEASURES

Precision: $P_{p,c} = \frac{|C_{p,c}|}{|R_{p,c}|}$

Pseudo-recall:

$$\alpha_{p,c} = \frac{|C_{p,c}|}{|C_{Pagico}|+|C_{aval}|}$$

Pseudo-F-measure:

$$\phi_{p,c} = 2 \times \frac{P_{p,c} \times \alpha_{p,c}}{P_{p,c}+\alpha_{p,c}}$$

Originality:

$$O_{p,c} = \sum_i^T \sum_j^{R_{p,c,i}} o(r_{p,c,i,j})$$

Creativity:

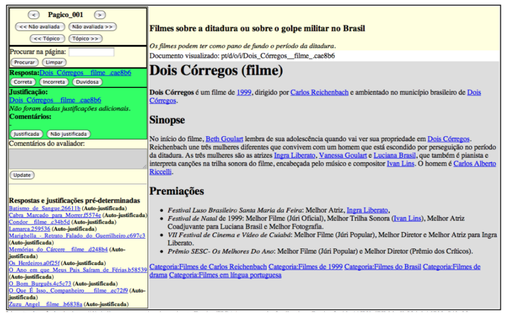
$$K_{p,c} = \sum_i^T \sum_j^{R_{p,c,i}} k(r_{p,c,i,j})$$

Final score: $M_{p,j} = |C_{p,c}| \times P_{c,j}$

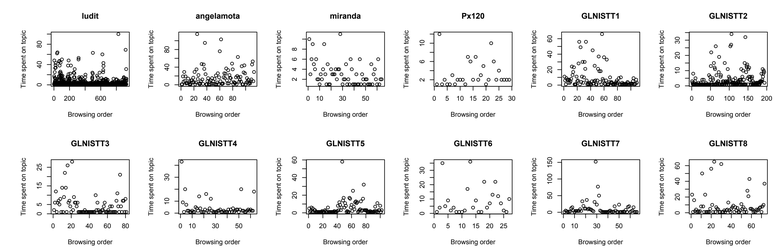
In addition to the measures used in GikiP and GikiCLEF, we chose to investigate originality and creativity, by weighing differently answers according to the number of participants who found them.

SIGA

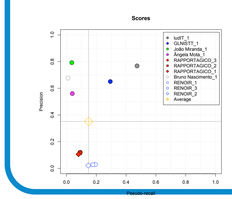
- Topic creation
- System run submission and testing
- Human participation interface
- Assessment interface
- Conflict resolution
- Pool browsing
- Scoring

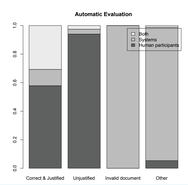


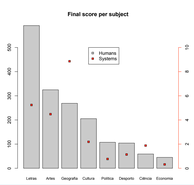
USER BROWSER BEHAVIOUR

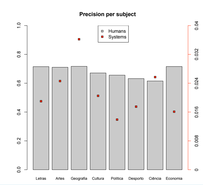


HUMANS VS. SYSTEMS





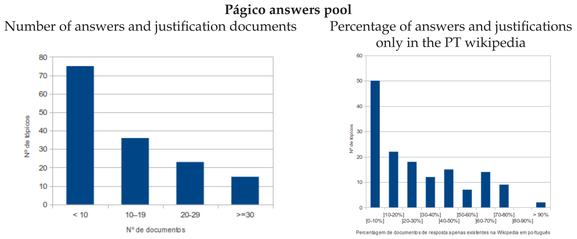




The most correct topics

	ID	Topic	Total	Hum	Sys	H & S
H	135	Aves de Angola	54	10	**44**	0
	19	Tribos indígenas que vivem na Amazônia.	115	56	**35**	24
	90	Filmes brasileiros premiados na categoria Montagem.	34	8	**19**	7
	13	Dinossauros carnívoros que habitaram o Brasil.	23	6	**12**	5
S	19	Tribos indígenas que vivem na Amazônia.	115	56	35	**24**
	62	Praias de Portugal boas para a prática de surf	30	5	6	**19**
	7	Guitarristas portugueses que também foram compositores.	34	17	0	**17**
	11	Filmes sobre o cangaço.	41	20	4	**17**

CARTOLA http://www.linguateca.pt/Cartola



REFERENCES

Diana Santos, Nuno Cardoso, Paula Carvalho, Iustin Dornescu, Sven Hartrumpf, Johannes Leveling, and Yvonne Skalban. 2009. GikiP at GeoCLEF 2008: Joining GIR and QA forces for querying Wikipedia. In C. Peters, T. Deselaers, N. Ferro, J. Gonzalo, G. J.F.Jones, M. Kurimo, T. Mandl, A. Pe nas, and V. Petras, ed., *Evaluating Systems for Multilingual and Multimodal Information Access 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers*, Springer, pp. 894-905.

Diana Santos and Luís Miguel Cabral. 2010. GikiCLEF : Expectations and lessons learned. In Carol Peters, Giorgio Di Nunzio, Mikko Kurimo, Thomas Mandl, Djamel Mostefa, Anselmo Peñas, and Giovanna Roda, ed., *Multilingual Information Access Evaluation, VOL I*, Springer, pp. 212-222.

Diana Santos, Cristina Mota, Cláudia Freitas, and Luís Costa. 2012. *Linguamática* **4**, number 1, special volume about Págico.

ACKNOWLEDGMENTS

Linguateca has throughout the years been jointly funded by the Portuguese Government, the European Union (FEDER and FSE), UMIC, FCCN and FCT. Págico was also supported by the Universities of Oslo, PUC-Rio, Coimbra and FCT grant SFRH/BPD/73011/2010.