C. Gurrin et al (Eds.): Advances in Information Retrieval
Proceedings of the 32nd European Conference on Information Retrieval, ECIR 2010
Milton Keynes, UK, pp. 672, ©Springer 2010.

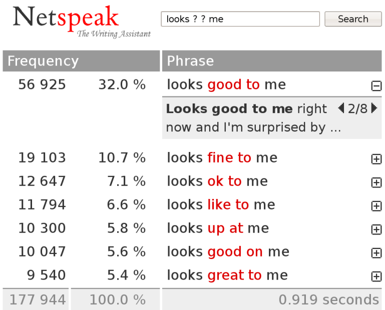
NETSPEAK—Assisting Writers in Choosing Words

Martin Potthast, Martin Trenkmann, and Benno Stein

Bauhaus-Universität Weimar, Germany
<first name>.<last name>@uni-weimar.de

NETSPEAK is a Web service which helps writers in finding alternative expressions for what they want to say.[1] It provides a large index of writing samples in the form of n-grams, $n \leq 5$, along with an efficient means to retrieve them by the use of wildcard queries. When in doubt about a phrasing, a user can get additional evidence by retrieving samples that match a given context. The figure below shows the results for a query where a user is interested in the two most frequently written words between "looks" and "me". The first two columns give an idea about the customariness of each result, and the user can select the one most appropriate for her sentence.

To provide a rich choice of writing samples we index the Google n-gram corpus which was compiled from a large portion of the English Web and which consists of more than 3 billion n-grams along with their occurrence frequencies [2]. We have developed a space-optimal inverted index based on minimal perfect hashing. The hash function maps the vocabulary V of the corpus to the storage positions of postlists. A hash function is perfect if it does not produce hash collisions for the key set V, and it is minimal if the number of storage positions required does not exceed $|V|$. The hash function is constructed with the CHD algorithm which produces a space overhead of $2.07 \times |V|$ bits [1]. Moreover, the index provides a top-k retrieval strategy to find the n-grams matching a query; details can be found in [3]. The table below shows selected performance data of our index. NETSPEAK is currently deployed on a cluster of 15 computers. In a load test the service was measured to process about 10 000 queries per second.



Frequency		Phrase	
56 925	32.0 %	looks good to me	⊟
		Looks good to me right ◀ 2/8 ▶ now and I'm surprised by ...	
19 103	10.7 %	looks fine to me	⊞
12 647	7.1 %	looks ok to me	⊞
11 794	6.6 %	looks like to me	⊞
10 300	5.8 %	looks up at me	⊞
10 047	5.6 %	looks good on me	⊞
9 540	5.4 %	looks great to me	⊞
177 944	100.0 %	0.919 seconds	

NETSPEAK *Benchmarks*	
Index size (compressed)	66.2 GB (31.8 GB)
Indexing time (single PC)	7:31 hours
Avg. retrieval time	0.164 seconds
Longest postlist ("**,**")	178 691 474 entries
Avg. postlist length	5334 entries
Index Scalability	
Index size	# postlists × postlist size
# postlists (key set size)	max. unsigned 32 bit integer
maximal postlist size	max. allowed file size

Bibliography

[1] D. Belazzougui, F.C. Botelho, and M. Dietzfelbinger. Hash, Displace, and Compress. *Proc. of ESA'09.*

[2] T. Brants and A. Franz. Web 1T 5-gram Version 1. Linguistic Data Consortium, 2006.

[3] B. Stein, M. Potthast, and M. Trenkmann. Retrieving Customary Web Language to Assist Writers. *Proc. of ECIR'10.*

[1] NETSPEAK is accessible at http://www.netspeak.cc