C. Gurrin et al (Eds.): Advances in Information Retrieval
Proceedings of the 32nd European Conference on Information Retrieval, ECIR 2010
Milton Keynes, UK, pp. 640-644, ©Springer 2010.

Cross-language High Similarity Search
Why no Sub-linear Time Bound can be Expected

Maik Anderka, Benno Stein, and Martin Potthast

Bauhaus University Weimar, Faculty of Media, 99421 Weimar, Germany
<first name>.<last name>@uni-weimar.de

Abstract This paper contributes to an important variant of cross-language information retrieval, called cross-language high similarity search. Given a collection D of documents and a query q in a language different from the language of D, the task is to retrieve highly similar documents with respect to q. Use cases for this task include cross-language plagiarism detection and translation search. The current line of research in cross-language high similarity search resorts to the comparison of q and the documents in D in a multilingual concept space—which, however, requires a linear scan of D. Monolingual high similarity search can be tackled in sub-linear time, either by fingerprinting or by "brute force n-gram indexing", as it is done by Web search engines. We argue that neither fingerprinting nor brute force n-gram indexing can be applied to tackle cross-language high similarity search, and that a linear scan is inevitable. Our findings are based on theoretical and empirical insights.

1 High Similarity Search

In the literature the task of high similarity search is also referred to as near-duplicate detection or nearest neighbor search. High similarity search techniques are applied in many applications such as for duplicate detection on the Web, text classification and clustering, plagiarism detection, or storage maintenance.

Without loss of generality we consider a document d represented under a bag of words model, as an m-dimensional term vector $\mathbf{d}$. The similarity between a query document q and a document d is quantified with a measure $\varphi(\mathbf{q}, \mathbf{d}) \in [0; 1]$, with 0 and 1 indicating no and maximum similarity respectively. φ may be the cosine similarity.

Definition 1 (High Similarity Search). *Given a query document q and a (very large) collection D of documents, the task of high similarity search is to retrieve a subset $D_q \subset D$, containing the most similar documents with respect to q:*

$$d \in D_q \;\; \Rightarrow \;\; \varphi(\mathbf{q}, \mathbf{d}) \geq 1 - \epsilon \tag{1}$$

D_q is called ϵ-neighborhood of q; a document $d \in D_q$ is called near-duplicate of q.

Since $\mathbf{q}$ and $\mathbf{d}$ are term-based representations, a document d is considered as near-duplicate of the document q if d and q share a very large part of their vocabulary. This syntactic definition of near-duplicate cannot be applied between two languages L and L', and cross-language near-duplicates need to be defined in a semantic manner. Consider for example a document d in language L that is a translation

of a document q' in language L'. Then similarity can be measured by a multilingual retrieval model that maps q' and d into a common, multilingual concept space. A few multilingual retrieval models exist, for a comparative overview see [6]. One of the most promising multilingual retrieval models is Cross-language Explicit Semantic Analysis, CL-ESA [1,6], which exploits a document-aligned comparable corpus such as Wikipedia in order to represent documents written in different languages in a common concept space. The cross-language similarity between q' and d is computed as cosine similarity $\varphi_{cos}(\mathbf{q}'_{clesa}, \mathbf{d}_{clesa})$ of the CL-ESA representations of q' and d.

Definition 2 (Cross-language High Similarity Search). *Given a query document q' in language L' and a (very large) collection D of documents in language L, the task of cross-language high similarity search is to retrieve a subset $D_{q'} \subset D$, containing the most similar documents with respect to q':*

$$d \in D_{q'} \;\Rightarrow\; \varphi_{cos}(\mathbf{q}'_{clesa}, \mathbf{d}_{clesa}) \geq 1 - \epsilon \tag{2}$$

$D_{q'}$ is called ϵ-neighborhood of q'; a document $d \in D_{q'}$ is called near-duplicate of q'.

To determine the value of ϵ we empirically analyzed the similarity values of near-duplicates in both settings: monolingual high similarity search and cross-language high similarity search. The left plot in Figure 1 shows the distribution of similarities between randomly selected English Wikipedia articles and their revisions—which serve as near-duplicates—computed as defined in (1). The right plot in Figure 1 shows the distribution of cross-language similarities (*i*) between randomly selected aligned English and German Wikipedia articles, i.e., the articles describe the same concept in its respective language, and (*ii*) between randomly selected aligned English and German documents from the JRC-Acquis corpus, which contains professional translations. In both cases the aligned documents are considered as cross-language near-duplicates, the respective cross-language similarities are computed as defined in (2). The analysis shows that the absolute similarity values of near-duplicates heavily differ in monolingual high similarity search and cross-language high similarity search. In the former a reasonable ϵ to detect the near-duplicates has to be very small (~ 0.15), whereas, in the latter a reasonable ϵ to detect cross-language near-duplicates has to be much higher (~ 0.5). One explanation for the relatively small similarity values of the cross-language near-duplicates is that the CL-ESA model is not able to operationalize the concept of "semantic similarity" entirely; and, it is still questionable if this is possible at all. However, even if the absolute similarity values of cross-language near-duplicates are relatively small, cross-language high similarity search is still possible since the average cross-language similarities between randomly selected documents—which are not aligned—of Wikipedia as well as the JRC-Acquis is about 0.1.

2 Linear Scan

A naive approach to high similarity search is a linear scan of the entire collection, i.e., calculating $\varphi(\mathbf{q}, \mathbf{d})$ or $\varphi_{cos}(\mathbf{q}'_{clesa}, \mathbf{d}_{clesa})$ for all $d \in D$. The retrieval time is $O(|D|)$, which is unfeasible for practical applications when D is very large, e.g. the World Wide Web. However, there are several approaches that try to speed up the pairwise similarity calculation in practice, e.g., by distributing the similarity computation based on MapReduce [4], or by using a specialized inverted index in combination with several heuristics

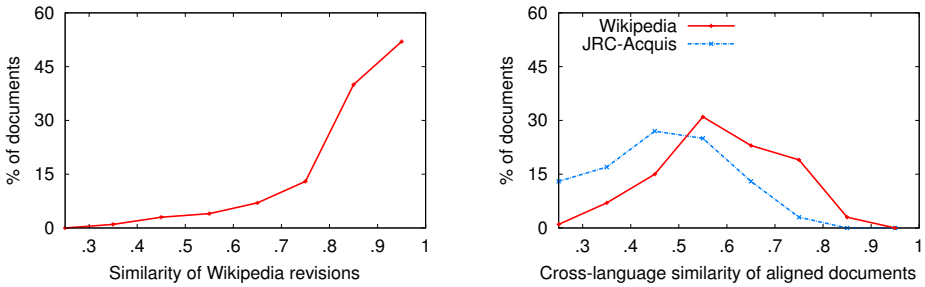

Figure 1. The left plot shows the distribution of similarities between English Wikipedia articles and their revisions, the right plot shows the distribution of cross-language similarities between aligned English and German documents from Wikipedia and the JRC-Acquis corpus.

to reduce the number of required multiplications [2]. In low-dimensional applications ($m < 10$) similarity search can be accelerated by means of space- or data-partitioning methods, like, grid-files, kd-trees, or R-trees. However, if the dimensionality is larger than 10—which is usual in practical applications, where the documents are represented as high dimensional feature vectors—these methods are outperformed by a simple linear scan [7].

3 Fingerprinting

Hash-based search or fingerprinting does not depend on the dimensionality of the feature vectors and allows for monolingual high similarity search in sub-linear retrieval time. Fingerprinting approaches simplify a continuous similarity relation to the binary concept "similar or not similar". A multi-valued similarity hash-function h_φ is used to map a feature vector $\mathbf{d}$ onto a small set of hash codes $F_d := h_\varphi(\mathbf{d})$, called fingerprint of d. Two documents q and d are considered as similar if their fingerprints share some hash code: $F_q \cap F_d \neq \emptyset \Rightarrow \varphi(\mathbf{q}, \mathbf{d}) \geq 1 - \epsilon$, with $0 < \epsilon \ll 1$. The mapping between all hash codes $C := \bigcup_{d \in D} F_d$ and documents with the same hash code can be organized as a hash table $\mathcal{T} : C \rightarrow \mathcal{P}(D)$. Based on $\mathcal{T}$ the set D_q can be constructed in $O(|D_q|)$ runtime as $D_q = \bigcup_{k \in F_q} \mathcal{T}(k)$. In most practical applications $O(|D_q|)$ is bound by a small constant since $|D_q| \ll |D|$; the cost of a hash table lookup is assessed with $O(1)$. Many fingerprinting approaches are described in the literature, which mainly differ in the design of h_φ. For a comparative overview see [5].

A similarity hash-function h_φ produces with a high probability a hash collision for two feature vectors $\mathbf{q}$, $\mathbf{d}$, iff $\varphi(\mathbf{q}, \mathbf{d}) \geq 1 - \epsilon$, with $0 < \epsilon \ll 1$. This is illustrated by an empirical analysis of different fingerprinting approaches in a monolingual high similarity search scenario, see Figure 2. All approaches achieve reasonable precision and recall at high similarities (~ 0.9). As shown in Section 1 the similarity values of cross-language near-duplicates are on average 0.5 (see Figure 1) where the recall of hash-based search drops dramatically. Hence, fingerprinting is not applicable to tackle cross-language high similarity search.

4 Brute Force Indexing

Web search engines solve the task of high similarity search very efficiently by indexing the collection D based on n-grams. Their “brute force n-gram indexing” strategy can

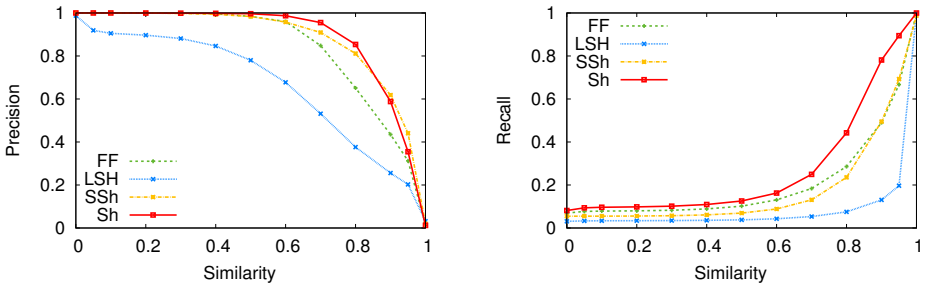

Figure 2. Precision and recall over similarity for fuzzy-fingerprinting, FF, locality-sensitive hashing, LSH, supershingling, SSh, and shingling, Sh [5].

be interpreted as a special case of fingerprinting if a query is considered as single n-gram with a reasonable large n, i.e. $n \in [5; 15]$. An example is the phrasal search functionality of a Web search engine. E.g. a Google query that is set into quotation marks is treated as a single n-gram. Consider a string-based hash function h, like MD5 or Rabin's hash function, that maps an n-gram onto a single hash code. The fingerprint F_d of a document d can be defined as $F_d = \bigcup_{c \in N_d} h(c)$, where N_d denotes the set of all n-grams of d. As described above, the mapping between hash codes and documents can be organized as a hash table $\mathcal{T}$. Since the query q is assumed to be a single n-gram, i.e. $|N_q| = 1$, the set D_q can be constructed as $D_q = \mathcal{T}(h(q))$. The runtime corresponds to a single hash table lookup, which is assessed with $O(1)$.

However, brute force n-gram indexing is not applicable to cross-language high similarity search since the hash codes $h(q')$ and $h(d)$ of a query q' in language L' and some document $d \in D$ in language L are not comparable.

5 Conclusion

For cross-language high similarity search no sub-linear time bound can be expected. We argued in this paper why—in contrast to monolingual high similarity search—neither fingerprinting nor brute force n-gram indexing can be used to model cross-language similarities that are close to 1. In our current research we use the LSH framework of Motwani to derive theoretical performance bounds for cross-language fingerprinting.

References

1. M. Anderka and B. Stein. The ESA Retrieval Model Revisited. In *Proc. of SIGIR'09*.
2. R. J. Bayardo, Y. Ma, and R. Srikant. Scaling Up All Pairs Similarity Search. In *Proc. of WWW'07*.
3. M. Datar, N. Immorlica, P. Indyk, and V. S. Mirrokni. Locality-Sensitive Hashing Scheme Based on p-Stable Distributions. In *Proc. of SCG'04*.
4. J. Lin. Brute Force and Indexed Approaches to Pairwise Document Similarity Comparisons with MapReduce. In *Proc. of SIGIR'09*.
5. M. Potthast and B. Stein. New Issues in Near-duplicate Detection. In *Data Analysis, Machine Learning and Applications*, 2008.
6. M. Potthast, B. Stein, and M. Anderka. A Wikipedia-Based Multilingual Retrieval Model. In *Proc. of ECIR'08*.
7. R. Weber, H.-J. Schek, and S. Blott. A Quantitative Analysis and Performance Study for Similarity-Search Methods in High-Dimensional Spaces. In *Proc. of VLDB'98*.