Light Stemming Approaches for the French, Portuguese, German and Hungarian Languages

Jacques Savoy
Institut interfacultaire d'informatique, University of Neuchatel,
Pierre-à-Mazel 7, 2000 Neuchâtel, Switzerland
Jacques.Savoy@unine.ch

ABSTRACT
This paper describes and evaluates various general stemming approaches for the French, Portuguese (Brazilian), German and Hungarian languages. Based on the CLEF test-collections, we demonstrate that light stemmers for the French, Portuguese and Hungarian languages perform well, and reasonably well for the German language. Variations in mean average precision among the different stemming approaches are also evaluated and sometimes they are found statistically significant.

Categories and Subject Descriptors
H.3.1 [**Content Analysis and Indexing**]: *Indexing methods; Linguistic processing*. H.3.3 [**Information Search and Retrieval**]: *Retrieval models*. H.3.4 [**Systems and Software**]: *Performance evaluation.*

General Terms
Algorithms, Measurement, Performance.

Keywords
Stemming for French, Portuguese, German, Hungarian; stemmer, natural language processing.

1. INTRODUCTION
Stemming is the conflation of word variants into a common stem, and in information retrieval (IR), applying stemmers to index documents (and requests) is assumed to be a good practice. Although *N*-gram [1] indexing would be an exception, we usually assume that stemming will efficiently conflate several different word variants into a common form. For example, when a query contains the word "horse," it seems reasonable to also retrieve documents containing the related word "horses." Properly designed stemming procedures should also reduce the size of inverted files, and also be helpful in other uses such as text data mining, natural language processing or gathering statistics on a document corpus.

In our approach to stemmer design, we begin by removing only inflectional suffixes. As such, singular and plural word forms (e.g., "dogs" and "dog") or feminine and masculine variants (e.g., "actress" and "actor") will conflate to the same root. Stemming schemes that remove only morphological inflections are termed "light" suffix-stripping algorithms. More sophisticated approaches exist for removing derivational suffixes (e.g., '-ment', '-ably', '-ship' in the English language). Stemming procedures [2], [3] ignore word meanings and tend to make errors, usually due to over-stemming (e.g., "organization" is reduced to "organ") or to under-stemming (e.g., "create" and "creation" do not conflate to the same root). Stemmers are usually designed to work with general text in any given language, yet most studies on their IR performance have involved English language stemmers only. Given the absence of evaluation studies for other European languages, this paper is intended to fill this gap.

SAC'06, April 23–27, 2006, Dijon, France.

2. RELATED WORK
Most stemming approaches are based on the target language's morphological rules (e.g., [2], [3]) and suffix removal is also controlled by quantitative restrictions (e.g., '-ing' is removed when the resulting stem has more than three letters as in "running," but not in "king") or qualitative restrictions (e.g., '-ize' is removed if the resulting stem does not end with "e" as in "seize"). Certain ad hoc spelling correction rules can also be applied to improve conflation accuracy (e.g., "running" gives "run" and not "runn"), particularly when phonetic rules are applied to facilitate easier pronunciation.

Another approach consults an online dictionary to obtain better conflation results [4], [6], while Xu & Croft [5] suggest a corpus-based approach that more closely reflects the language at hand rather than its grammar. Few stemming procedures[1] have been suggested for other European languages than English, and those schemes available usually apply to the more popular languages and rely on a dictionary [6] or a deeper morphological analysis [7].

Following a performance analysis of various English stemmers, Harman [8] showed there were no statistically significant improvements between them. A query-by-query analysis did however reveal that stemming affected performance, even though the number of queries showing improvement was almost equal to the number of queries showing decreased performance. Other studies generally conclude that the use of a stemmer shows a modest improvement, and the difference with an approach ignoring stemming is not always statistically significant.

It was also surprising to note that during recent CLEF evaluation campaigns,[2] only a few stemmers were suggested and compared. For example, Di Nunzio *et al*. [9] showed that for statistical stemmers the relative retrieval performance may

[1] Freely available at the Web site http://snowball.tartarus.org/ or http://www.unine.ch/info/clef/

[2] See the Web page http://clef.iei.pi.cnr.it/

vary for each of the five studied languages. This means that such an approach may work well for one language yet work poorly for another. When compared to statistical stemmers, Porter's stemmers seem to work slightly better. For German, Braschler & Ripplinger [10] showed that for short queries stemming may enhance mean average precision by 23%, compared to 11% for longer queries. Finally, Tomlinson [11] evaluated the differences between Porter's stemmer [3] and the lexical stemmer (based on a dictionary of the corresponding language). For Finnish and German, he found that the lexical stemmer tended to produce statistically better results, while for seven other languages performance differences were insignificant.

Based on these facts, the rest of this paper will address the following questions: 1) Does stemming affect IR performance for European languages other than English? 2) For these languages, are light stemming approaches less effective than more complex suffix-stripping algorithms?

3. TEST-COLLECTIONS

The corpora we used are from the CLEF'05 evaluation campaign, and consist of newspaper and news agency articles. The German collection is part of the GIRT corpora and is composed of bibliographic records extracted from various sources in the social sciences. A typical record in this German corpus consists of a title, an abstract and a set of manually assigned descriptors. See Kluck [12] for a more complete description of this corpus.

Table 1. Some statistics from our test-collections (CLEF)

	French	Portuguese	Hungarian	German
Size	487 MB	564 MB	105 MB	326 MB
# docs	177,452	210,734	49,530	151,319
mean terms	178	212.9	142.1	89.6
# queries	50	50	50	50
# rel. doc/q	50.7	58.1	18.8	86.9

As shown in Table 1, both the French and Portuguese corpus have roughly the same size (487 MB vs. 564 MB), while the German ranks second and the Hungarian third, both in size (105 MB) and in number of documents (49,530). During the indexing process, we only retained those logical sections allowed by CLEF evaluation campaigns, meaning a priori that all pertinent sections were used to build document representatives. For the German collection, we applied a decompounding procedure [13], retaining both the compounds and their component words in document or topic representations. Compound words (e.g., newspaper, courtroom) are widely used and with variants in German and thus they lead to more difficulties than they those of the English language; for example research project is "Forschungsprojekt," combining "Forschung" + s + "Projekt." Finally, accents are removed, even though this process may accidentally conflated words with different meanings into the same form (e.g., in French the word "tâche" (task) and "tache" (mark, spot)).

Based on the TREC model, each topic was structured into three logical sections comprising a brief title, a one-sentence description, and a narrative part specifying the relevance assessment criteria. In this study, we used the shortest query formulation in order to reflect a more realistic search context. Based on the topic title-only, queries had a mean size of 2.8 search terms for the French collection, 2.6 for the Portuguese, 2.2 for the Hungarian and 1.7 for the German.

The available topics covered various subjects (e.g., "Money Laundering", or "Lottery Winnings") and included both regional ("Golden Bear") and international coverage ("Anti-abortion Movements"). The same set of queries was used for the French, Portuguese and Hungarian collections, while the German corpus was searched using a different set of 50 topics (e.g., "Religion and Politics" or "Electoral Behaviour").

As shown in Table 1, the number of relevant items per query for the French and Portuguese collection has a relatively similar mean value (50.7 and 58.1 respectively), and lower for the Hungarian corpus (18.8). The size of this collection however was only one quarter the size of the French corpus. The mean number of relevant articles per request for the German test-collection was clearly higher, having a mean value of 86.9.

4. IR MODELS

To ground our findings on solid foundations and to obtain a broader view of the relative merit of the various retrieval models, we used nine vector-space and two probabilistic IR models to evaluate the various stemming approaches. First we adopted a binary indexing scheme in which each document (or request) was represented by a set of keywords, without any weight. To measure similarities between documents and requests we computed the inner product (model denoted "doc=bnn, query=bnn" or "bnn-bnn"). We could take term occurrence frequency into account (or tf) with the corresponding retrieval model being denoted as "nnn-nnn". We could also account for their inverse document frequency (or idf) and also normalize each indexing weight using different weighting schemes, as is described in [13].

Other variants may be created, especially given that the occurrence of a particular term in a document is a rare event. Thus, we may assign more importance to the first occurrence of a word, as compared to any successive, repeating occurrences. Therefore, the tf component would be computed as the ln(tf) + 1.0 (model denoted "ltc-ltc") or as 0.5 + 0.5 · [tf / max tf in a document]. Different weighting formulae may of course be used for documents and requests, leading to other different weighting combinations. We might also consider that a term's presence in a shorter document provides stronger evidence than it does in a longer document, leading to more complex IR models; for example, the IR model denoted by "doc=Lnu" [14], "doc=dtu" [15].

In addition to these vector-space schemes, we also considered probabilistic models such as the Okapi model [16]. As a second probabilistic approach, we implemented the Prosit approach [17], based on combining two information measures that are formulated as follows:

$$w_{ij} = Inf^1_{ij} \cdot Inf^2_{ij} = (1 - Prob^1_{ij}) \cdot -\log_2[Prob^2_{ij}]$$
$$Prob^1_{ij} = tfn_{ij} / (tfn_{ij} + 1),\ tfn_{ij} = tf_{ij} \cdot \log_2[1 + ((C \cdot mean\ dl) / l_i)]$$
$$Prob^2_{ij} = [1 / (1+\square_j)] \cdot [\square_j / (1+\square_j)]^{tfn_{ij}} \quad \text{with } \square_j = tc_j / n$$

where w_{ij} indicates the indexing weight attached to term t_j in document D_i, l_i the number of indexing terms included in the representation of D_i, where tc_j represents the number of occurrences of term t_j in the collection, n the number of documents in the corpus, and C and *mean dl* are constants.

5. STEMMING STRATEGIES

Over the last few years we designed light stemming procedures for the various European languages. We believe it important to develop a simple approach, one that does not require a dictionary or any other sophisticated data structures or processing. We also believe that effective stemming should focus mainly on nouns and adjectives, thus ignoring various verb forms (although past participles could be an exception to this rule). In this vein, our stemming approach tried to remove the morphological variations associated with number (singular vs. plural), gender (masculine or feminine), and various grammatical cases (nominative, accusative, ablative, etc.). In verbal forms we ignored variations which are usually fairly numerous, while for the adjectives we did not attempt to remove comparative and superlative suffixes.

An analysis of the grammar of any given language however usually reveals numerous inflectional rules, some of which are used for only one or a few words (e.g., "box" and "boxes" or "mouse" and "mice" in English). For those languages having a more complex morphology than English, we could develop a simple stemmer, based on just a few but frequently used rules. For the French language, such a stemming approach (label "S-stemmer") would be based on seven rules, as depicted in Table 2. For example, the word "chevaux" (horses) is reduced to "cheval" (horse) and the words "baronnes" (baronesses), "barons" and "baron" are reduced to the same stem "baron".

As a variant for the French language, we could suggest removing other inflections (basic parts shown in Table 2) and also certain derivational suffixes. Labeled "UniNE" in our experiments, this stemming is composed of 27 rules.

Table 2. Minimal S-stemmer for French language

For words of six or more letters if final is '-aux' then replace '-aux' by '-al', return; if the final letter is '-x' then remove '-x', return; if final letter is '-s' then remove '-s'; if final letter is '-r' then remove '-r'; if final letter is '-e' then remove '-e'; if final letter is '-é' then remove '-é'; if final two letters are the same, remove the final letter, return.

For Portuguese, our stemmer would try to remove inflections attached to both nouns and adjectives, based on rules for the plural form (10 rules) and feminine form (13 rules). In Portuguese as in English the plural form is usually obtained by adding an '-s' (e.g., "amigo" and "amigos" (friend)). This suffix is also used for adjectives. There are of course various exceptions to the general rule (e.g., "mar" and "mares" (sea), "fuzil" and "fuzis" (gun), and for the adjective "fácil" (easy), its plural form is "fáceis"). The feminine form is usually obtained by replacing the final '-o' by an '-a' (e.g., "americano" and "americana"), but there are various exceptions to be taken into account (e.g., "inglês" (British) becomes "inglesa" in the feminine, "leão" (lion) becomes "leoa" and "professor" gives "professora").

For German our suggested stemmer would incorporate 11 rules to remove both plural forms and grammatical case endings (e.g., those usually used to indicate the genitive case by employing an '-s' or '-es' as in "Staates" (of the state), "Mannes" (of the man)). In German the plural form is denoted using a variety of endings such as '-en' (e.g., "Motor" and "Motoren" (engine)), '-er', '-e' (e.g., "Jahr" and "Jahre" (year)) or '-n' (e.g., "Name" and "Namen" (name)). Plural forms also use diacritic characters (e.g., "Apfel" (apple) becomes "Äpfel" in its plural form) or in conjunction with a suffix (e.g., "Haus" and "Häuser" (house)). Also frequently used are the suffixes '-en' or '-n' to indicate grammatical cases or for adjectives (e.g., "… einen guten Mann" (a good man) in the accusative singular form).

As with the Finnish language, Hungarian makes use of a greater number of grammatical cases (usually 18) than German (four cases). Each case has its own unambiguous suffix; e.g. the noun "house" ("hàz" in nominative) may appear as "hàzat" (accusative case), "hàzakat" (accusative plural case, as in "(I see) the houses"), "hàzamat" ("… my house") or "hàzamait" ("… my houses"). In this language the following general construction is used for nouns: 'stem' 'plural' 'possessive marker' 'case' as in 'hàz' + 'am' + 'at' (in which the letter 'a' is introduced to facilitate better pronunciation because "hàzmt" could be difficult to pronounce). Our suggested "UniNE" stemmer is based on two rules for plural removal, 17 rules for removing various possessive suffixes and 21 rules for removing case markers. In a lighter stemming procedure, we would ignore the possessive marker (under the assumption that such suffixes are infrequently used and in an effort to reduce the number of conflation errors). Thus, in order to automatically remove the most frequent cases we would apply only 13 rules.

Compared to the 260 rules used in [2] or the 60 in [3] for the English language, for those languages having more a morphology more complex than English, the stemmers suggested could be viewed as light versions. These stemmers are freely available at http://www.unine.ch/info/clef/. As an alternative to our light stemmers, we might also employ a more aggressive stemmer, taken from those found within Porter's[3] family (available for the French, Portuguese and German languages).

6. EVALUATION

To measure retrieval performance, we adopted a non-interpolated mean average precision (MAP) computed by the TREC_EVAL program. To statistically determine whether or not a given search strategy would be better than another, we applied the bootstrap methodology [18]. In our statistical testing, the null hypothesis H_0 states that both retrieval schemes produce similar performance. Such a null hypothesis would be accepted if two retrieval schemes returned statistically similar means, otherwise it would be rejected. Thus in the tables included in this paper, we underline any statistically significant differences resulting from a two-sided non-parametric bootstrap test, and based on the MAP difference (significance level 5%).

6.1 IR Models Evaluation

Based on this methodology, Table 3 depicts the MAP for the French or Portuguese collections, using different stemming approaches. The same information is given in Table 4 for the Hungarian and German corpora. In these tables, the best performance under a given condition is shown in bold, and this will be used as the baseline for statistical testing. These experiments show that the Okapi probabilistic model usually produces the best retrieval performance across the different

[3] Freely available at the Web site http://snowball.tartarus.org/

Table 3. MAP of various IR models applying different stemming strategies (French & Portuguese collections)

	Mean average precision						
IR Model \ Stemmer	French none	French UniNE	French S-stemmer	French Porter	Portuguese none	Portuguese UniNE	Portuguese Porter
doc=Okapi, query=npn	**0.2260**	**0.3045**	**0.2858**	**0.2978**	**0.2238**	**0.2873**	**0.2610**
Prosit	0.2125	0.2918	0.2739	0.2878	0.2182	0.2755	0.2502
doc=Lnu, query=ltc	0.2112	0.2933	0.2717	0.2808	0.1989	0.2611	0.2296
doc=dtu, query=dtn	0.2062	0.2780	0.2611	0.2758	0.2096	0.2571	0.2189
doc=atn, query=ntc	0.2088	0.2755	0.2603	0.2695	0.2049	0.2458	0.2128
doc=ltn, query=ntc	0.1945	0.2466	0.2402	0.2371	0.1758	0.2149	0.1831
doc=lnc, query=ltc	0.1545	0.2233	0.2080	0.2131	0.1519	0.1811	0.1607
doc=ltc, query=ltc	0.1461	0.1975	0.1879	0.1922	0.1433	0.1625	0.1415
doc=ntc, query=ntc	0.1462	0.1918	0.1807	0.1758	0.1344	0.1553	0.1422
doc=bnn, query=bnn	0.1013	0.1153	0.1179	0.1017	0.1134	0.1309	0.1187
doc=nnn, query=nnn	0.0954	0.1013	0.1005	0.0894	0.0710	0.0630	0.0608

Table 4. MAP of various IR models applying different stemming strategies (Hungarian & German corpora)

	Mean average precision					
IR Model \ Stemmer	Hungarian none	Hungarian light	Hungarian UniNE	German none	German UniNE	German Porter
doc=Okapi, query=npn	0.1957	**0.2988**	**0.3076**	**0.3552**	**0.3931**	**0.4058**
Prosit	0.1883	0.2905	0.2964	0.3464	0.3805	0.3934
doc=Lnu, query=ltc	0.1887	0.2913	0.2868	0.3357	0.3638	0.3793
doc=dtu, query=dtn	**0.1980**	0.2857	0.2900	0.3357	0.3671	0.3826
doc=atn, query=ntc	0.1794	0.2651	0.2755	0.3381	0.3653	0.3789
doc=ltn, query=ntc	0.1919	0.2556	0.2567	0.3184	0.3421	0.3573
doc=lnc, query=ltc	0.1616	0.2188	0.2153	0.2757	0.2983	0.3032
doc=ltc, query=ltc	0.1675	0.2207	0.2183	0.2575	0.2773	0.2891
doc=ntc, query=ntc	0.1713	0.2162	0.2079	0.2510	0.2649	0.2759
doc=bnn, query=bnn	0.1338	0.1748	0.1782	0.2430	0.2552	0.2637
doc=nnn, query=nnn	0.1326	0.1348	0.1256	0.1381	0.1419	0.1462

languages (an exception to this finding is the Hungarian corpus without stemming, where the "dtu-dtn" approach (0.1980) produces a better MAP than the Okapi model (0.1957), this difference is however not statistically significant). For the French, Portuguese and German corpora however, differences between the Okapi model and other IR models are statistically significant.

6.2 Nonstemming vs. Stemming

In this section we would like to verify whether or not a stemming procedure might statistically improve MAP (and we will ignore both "bnn-bnn" and "nnn-nnn" models producing very poor retrieval effectiveness results). Retrieval performances without stemming will serve as the baseline (MAP depicted under the label "none" in Tables 3 and 4). For the French collection, all three stemming approaches performed statistically better than the baseline "none" for the nine IR models. After averaging percentage enhancement across these nine models, we found an average increase of 35% when using the UniNE stemmer, 30.5% with Porter's scheme, and 27.3% for the "S-stemmer".

For the Portuguese and German corpora, we found similar conclusions; with the two stemming procedures always performing statistically better than those done without stemming. When computing the MAP percentage differences across the nine IR models, we found that the UniNE stemmer improved the MAP by 22% on average for the Portuguese collection and by 8.4% for the German corpus. Using the same baseline, Porter's stemmer improved the MAP by 7.7% on average for the Portuguese collection, and by 12.4% for the German corpus.

For the Hungarian corpus, the two stemming approaches improved the MAP when compared to an approach not using stemming (on average by 42.8% for UniNE stemmer, and 42.2% for the light stemming scheme). Both stemmers did however statistically improve the MAP when compared to an indexing scheme that ignored stemming.

6.3 Comparing Different Stemmers

It is assumed that stemming usually improves the retrieval performance (even though performance differences are not always statistically significant) and that the different stemmers tend to produce similar results. To investigate this latest issue we compared the retrieval effectiveness produced by the various stemmers.

For the French collection and taking the "S-stemmer" retrieval performance as a baseline, Porter's stemmer improved by 2.5% on average (computed from the nine best performing IR models), although these differences are not statistically significant. For the UniNE stemmer, average enhancement was 6% (across nine IR models), a statistically significant difference for only the Okapi, Prosit, and "dtu-dtn" IR schemes. While the performance difference between Porter

and UniNE always favors the second (+3.5% in average), these variations are not however statistically significant.

For the Portuguese language, the situation is relatively similar. Using the UniNE stemmer as a baseline for the 9 IR models, Porter's approach gives lower MAP (-11.8% in average over nine IR models). Moreover, for 5 IR models, the difference is also statistically significant. Thus for both the French and the Portuguese languages, different stemmers may provide IR performances that could be statistically different. Moreover, for these languages at least a light stemming approach seemed to be more effective than a stemming approach that tried to remove some derivational suffixes.

For the German corpus, Porter's stemmer provided better retrieval performance than did the UniNE scheme (average difference of 3.7% over nine IR models). The difference between these two stemming schemes however was never statistically significant.

Finally for the Hungarian corpus, the difference between the two suggested stemming methods is very small (0.3% on average over nine IR models), and not statistically significant.

7. CONCLUSION

In this paper, upon analyzing four different languages and various stemming approaches, we demonstrated that the Okapi probabilistic model produces the best retrieval performance. Moreover, the difference between the MAP and other IR models is statistically significant for the French, Portuguese and German corpora.

Empirical evidence clearly shows that a stemming procedure improves retrieval effectiveness when applied to European languages belonging to either the Latin (French, Portuguese), Germanic (German) or Finno-Ugrian (Hungarian) families. From a statistical point of view, the difference in retrieval performance for these four languages is significant for the best nine performing IR models.

From comparing different stemming strategies, it seems a light stemming approach produces better MAP than does a more aggressive stemmer, and for some IR models, the difference between these two stemming schemes could be statistically significant and in favor of a light stemming solution. For the German and the Hungarian languages, the performance difference between the stemmers is not statistically significant.

ACKNOWLEDGMENTS

This research was supported in part by the Swiss National Science Foundation under Grants #21-66 742.01 and #200020-103420

8. REFERENCES

[1] McNamee, P., and Mayfield, J. Character *N*-gram tokenization for European language text retrieval. *IR Journal*, 7(1-2), 2004, 73-97.

[2] Lovins, J.B. Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, 11(1), 1968, 22-31.

[3] Porter, M.F. An algorithm for suffix stripping. *Program*, 14(3), 1980, 130-137.

[4] Krovetz, R. Viewing morphology as an inference process. In *Proceedings of the ACM-SIGIR*. Pittsburgh, PA, 1993, 191-202.

[5] Xu, J., and Croft, B. Corpus-based stemming using cooccurrence of word variants. *ACM-TOIS*, 16(1), 1998, 61-81.

[6] Savoy, J. Stemming of French words based on grammatical category. *JASIS*, 44(1), 1993, 1-9.

[7] Korenius, T., Laurikkala, J., Järvelin, K., and Juhola, M. Stemming and lemmatization in the clustering of Finnish text documents. In *Proceedings of the ACM-CIKM*. Washington, DC, 2004, 625-633.

[8] Harman, D. How effective is suffixing? *JASIS*, 42(1), 1991, 7-15.

[9] Di Nunzio, G.M., Ferro, N., Melucci, M., and Orio, N. Experiments to evaluate probabilistic models for automatic stemmer generation and query word translation. In *Comparative Evaluation of Multilingual Information Access Systems*. LNCS #3237, Springer, Berlin, 2004, 220-235.

[10] Braschler, M., and Ripplinger, B. How effective is stemming and decompounding for German text retrieval? *IR Journal*, 7(3-4), 2004, 291-316.

[11] Tomlinson, S. Lexical and algorithmic stemming compared for 9 European languages with Humminbird SearchServer™ at CLEF 2003. In *Comparative Evaluation of Multilingual Information Access Systems*. LNCS #3237, Springer-Verlag, Berlin, 2004, 286-300.

[12] Kluck, M. The GIRT data in the evaluation of CLIR systems – from 1997 until 2003. In *Comparative Evaluation of Multilingual Information Access Systems*. LNCS #3237, Springer, Berlin, 2004, 376-390.

[13] Savoy, J. Report on CLEF-2003 monolingual tracks. In *Comparative Evaluation of Multilingual Information Access Systems*. LNCS #3237, Springer, Berlin, 2004, 322-336.

[14] Buckley, C., Singhal, A., Mitra, M., and Salton, G. New retrieval approaches using SMART. In *Proceedings of TREC-4*. Gaithersburg, MA, 1996, 25-48.

[15] Singhal, A., Choi, J., Hindle, D., Lewis, D.D. & Pereira, F. (1999). AT&T at TREC-7. In *Proceedings TREC-7*, Gaithersburg, MA, 1999, 239-251.

[16] Robertson, S.E., Walker, S., and Beaulieu, M. Experimentation as a way of life: Okapi at TREC. *IP&M*, 36(1), 2000, 95-108.

[17] Amati, G., and van Rijsbergen, C.J. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM-TOIS*, 20(4), 2002, 357-389.

[18] Savoy, J. Statistical inference in retrieval effectiveness evaluation. *Information Processing &Management*, 33(4), 1997, 495-512.