# SENSITIVITY ANALYSIS OF INFORMATION RETRIEVAL METRICS

Marina Marjanović-Jakovljević

Department of Computer Engineering, Singidunum University, Belgrade, Serbia

## ABSTRACT

*Average Precision, Recall and Precision are the main metrics of Information Retrieval (IR) systems performance. Using Mathematical and empirical analysis, in this paper, we show the properties of those metrics. Mathematically, it is demonstrated that all those parameters are very sensitive to relevance judgment which is not usually very reliable. We show that position shifting downwards of the relevant document within the ranked list is followed by Average Precision decreasing. The variation of Average Precision parameter value is highly present in the positions 1 to 10, while from the 10th position on, this variation is negligible. In addition, we try to estimate the regularity of the Average Precision value changes, when we assume that we are switching the arbitrary number of relevance judgments within the existing ranked list, from non-relevant to relevant. Empirically, it is shown hat 6 relevant documents at the end of the 20 document list, have approximately same Average Precision value as a single relevant document at the beginning of this list, while Recall and Precision values increase linearly, regardless of the document position in the list. Also, we show that in the case of Serbian-to-English human translation query followed by English-to-Serbian machine translation, relevance judgment is significantly changed and therefore, all the parameters for measuring the IR system performance are also subject to change.*

## KEYWORDS

*Information Retrieval (IR) Systems, query, Ranking, Precision, Average Precision*

## 1. INTRODUCTION

Nowadays, following the constant development of information technologies, a significant amount of information exists and is available to everyone. It has been estimated that there is 55.5 % of documents in English on the Web. However, the English language is not the native language to 71.5 % of users [1]. On the other hand, the availability of the Internet has created wide opportunities for transcribing and translating the works of others, which seriously threatens the system of social values [2]. These facts create the necessity for Information Retrieval (IR) system which is capable of effective evaluation studies Cross Language IR, especially between English and other languages. Therefore, it is of essential importance to evaluate the effectiveness of the IR system. To assess the efficiency, it is necessary to establish suitable success criteria that can be measured in some way. Evaluation is important for designing, development and maintenance of IR system efficiency.

These may include performance assessment of IR system. Based on existing literature, it can be concluded that the evaluation of IR systems is the subject of research in the last 50 years [3]. This is due to the fact that feasibility study includes parameter such as customer satisfaction, laboratory results, and the results of operational tests, as it was discussed in [4] and [5].

The traditional way of measuring the effectiveness of IR system consists of a user assessment about relevance of particular document for a given query. This approach was influenced by Information Retrieval Community for IR algorithm development and TREC Conferences held in

DOI:10.5121/ijcsit.2015.7404

the USA, the third-TREC3 [6], the fourth-TREC4 [7] and the sixth-TREC-6 [8]. The algorithms developed at these events were based on the measurement of efficiency in a controlled experimental environment. The effectiveness of the IR system is estimated based on the speed of returning results as well as the necessary memory required to store the index. Measuring the effectiveness and efficiency of IR is usually carried out in the laboratory with minimal involvement of users and was based on the assessment of the results of completed algorithms. Robertson in [5] indicates that customer rating is also necessary in this evaluation in order to qualitatively estimate IR system performance. This user-oriented approach implies a state of users and their interaction with the IR system.

In practice, it is common to use different evaluation approaches that will be used during the development of IR systems, such as the use of the test collection for development, optimization algorithms for search, and laboratory experiments that involve the interaction of users who are involved in improving the user interface.

In Section II, we present a general IR system model. In Section III, we discuss measurement of the efficiency of the IR test collections that we use for the purpose of this paper as well as the parameters used to measure the quality of IR system. In this section, using Mathematical analysis, we show some interesting properties of those parameters. In Chapter IV, the criterion for the evaluation of relevant and irrelevant documents and the valuation methodology for the purpose experiment are presented. Finally, Section V concludes the results of this paper and gives some ideas about work.

## 2. IR SYSTEM MODEL

According to the model described in [9], in Figure 1, Flowchart system model is shown. The set of queries belongs to one reference corpus, while the set of original documents can belong to the other reference corpus. In the next stage, query document is divided into subdocuments and translated into the language of the documents collection. Furthermore, in the stage of Pre-processing, in order to shorten the time of analysis and make better IR system performance, the processes of stemming, stopword removal and Part of Speech (POS) tagging should be applied. Since the queries documents for experiments described in this paper are written in Serbian language, pre-processing should be adjusted to this language.

The classic stemmer for the Serbian language, where suffix-stripping is performed, is shown in [10] and [11].

The tagger for the Serbian language is described in [12], where the strategy to unify each separate case under a general rule was developed.

On the other side, the document from collection is submitted to indexing, where it receives its identification. Following the identification and Pre-processing stage, both documents are included in the Information Retrieval system adapted to the Serbian language in order to deal with Serbian alphabet and whose task is to estimate the degree to which documents in the collection reflect the information in a user query. Two methods for encoding Serbian characters are recommended in [13]:

ISO-8859-5, Cyrillic-based and suited for Eastern European languages (Bulgarian, Byelorussian, Macedonian, Russian, Serbian, and Ukrainian) and ISO-8859-2, suited for European languages (Albanian, Croatian, Czech, English, German, Hungarian, Latin, Polish, Romanian, Slovak, Slovenian, and Serbian).

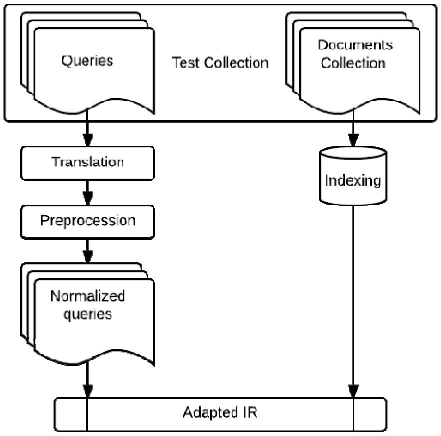

Figure 1. System Model Flowchart

## 3. Evaluation of IR System

For the measurement of ad-hoc information of search efficiency in a standard way, we must have a test collection consisting of three things:

A document collection, a test suite of information needs (expressible as queries), and relevance judgment for each query document pair.

As far as the information needs of users, a document in the collection should give a classification whether it is relevant or not. This decision is of great importance when the parameters for a quality measure of IR systems are calculated.

### 3.1. Test collection

In order to measure the efficiency of IR systems, it is necessary to have the appropriate test collection. There are collections such as TREC-3 [6], TREC-4 [7] for the Spanish language, TREC-6 [8], for Cross Language French and English, ECLAPA [9] for Cross Language Portuguese and English. In [14], testing for query collection written in Serbian language was done by EBART 3.

However, many IR systems contain different parameters that can be adjusted to tune its performance. Therefore, it is not correct to report results on a test collection which was obtained by tuning these parameters in order to maximize performance on one particular set of queries rather than for a random sample of queries. In this respect, in order to tune the parameters of the system, the correct procedure is to have more than one test collections.

### 3.2. IR System evaluation measures properties

There are several measures for the system operation for information search. The measures are based on a collection of documents and queries in which the relevance of the given documents is known. All the usual measures which are described here assume a binary relevance: the document is either relevant or non-relevant. In practice, queries can be badly placed and there may be different shades of relevance.

Evaluation of the quality of IR system is based on the calculation of Precision, Average Precision and Recall parameters [4].

Precision parameter represents the ratio of the number of retrieved relevant documents ( $N_{rel\_ret}$ ) and the total number of ranked documents ( $N_{ranked\_doc.}$ ), presented as result of the IR system.

$$\Pr ecision = \frac{N_{rel\_ret}}{N_{ranked\_doc.}} \tag{1}$$

Recall represents the ratio of the number of relevant retrieved documents ( $N_{rel\_ret}$ ) and the total number of relevant documents within the document collection ( $N_{rel}$ ).

$$\mathrm{Re}\,call = \frac{N_{rel\_ret}}{N_{rel}} \tag{2}$$

In the case that we want Average over queries, more complicated joint measure is required as the Average Precision. Unfortunately, Recall and Precision have been rarely used in recent retrieval experiments since all retrieved documents to be ranked and checked for computing results of these indicators requires large computing time.

For a single information need, in [15], Average Precision is defined as the mean of the Precision scores obtained after each relevant document is retrieved, using zero as the Precision for relevant documents that are not retrieved.

In order to better understand those metrics and their sensitivity to the relevance judgment, we will suppose that we need to evaluate IR system by using a test collection which consists of a set of queries, set of documents and relevance judgments results whether the document from collection is relevant or not.

For the sake of simplicity, let $x_k$ be a variable that represents binary relevance judgments, i.e.

$$x_k = \begin{cases} 1 \;\; if \;\; k^{th} \;\; document \; is \; relevant \\ 0 \;\; if \;\; k^{th} \;\; document \; is \; not \; relevant \end{cases} \tag{3}$$

In order to calculate Precision, we will observe the $N_{ranked\_doc.}$ ranked documents in the set of documents. Therefore, from (1), Precision after $N_{ranked\_doc.}$ ranked documents retrieved can be computed as:

$$\Pr ecision_N = \frac{1}{N_{ranked\_doc.}} \sum_{k=1}^{N_{ranked\_doc.}} x_k \tag{4}$$

On the other side, using the same notations and considering definition of Recall parameter from (2), for $N_{ranked\_doc.}$ top ranked documents, the value of Recall is computed such that

$$\mathrm{Re}\,call_k = \frac{1}{N_{rel}} \sum_{k=1}^{N_{ranked\_doc.}} x_k \tag{5}$$

Therefore, using our notations, Average Precision can be expressed as

$$Average\,\mathrm{Pr}\,ecision\,(i) = \frac{1}{N_{rel\_ret}} \sum_{i=1}^{N_{ranked\_doc.}} x_i \,\mathrm{Pr}\,ecision_i = \frac{1}{N_{rel\_ret}} \sum_{i=1}^{N_{ranked\_doc.}} \frac{x_i}{i} \sum_{k=1}^{i} x_k \tag{6}$$

From the previous expressions, it is clear that the value of Average Precision depends on the number of relevant documents retrieved and the position of relevant documents within the ranked list. Supposing that the number of relevant documents in collection and the number of documents in the list have constant value and that we have only one relevant document $N_{rel\_ret}$=1 on the ith position within the document set ($1 \le i \le N_{ranked\_doc.}$), Average Precision parameter becomes:

$$Average\,\mathrm{Pr}\,ecision\,(i) = \frac{1}{i} \tag{7}$$

While Recall parameter becomes:

$$\mathrm{Re}\,call = \frac{1}{N_{rel}} = const. \tag{8}$$

Therefore, the difference in value of Average Precision when the relevant document is located on two adjacent positions within the ranked list is

$$dAverage\,\mathrm{Pr}\,ecision\,(i) = -\frac{1}{i \cdot (i+1)} \tag{9}$$

Those two functions are shown in Figure 2. Figure 2. Illustrates that the value of Average Precision and $\left| dAverage\,\mathrm{Pr}\,ecision\,(i) \right|$ are negligible for $i \ge 10$.

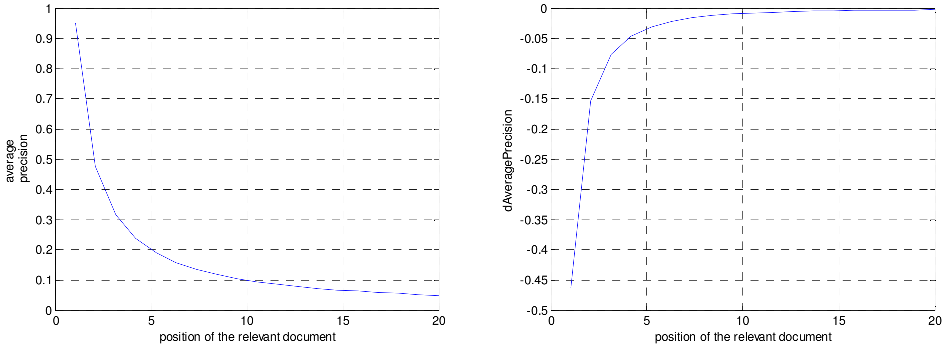

Figure 2. Average Precision (on the left one), dAveragePrecision (on the right one) vs. position of the relevant document: $N_{rel\_ret}$=1;

In a further discussion, we will try to estimate the regularity of change of the Average Precision value when we assume that we switch the relevance judgments of the documents from the ranked list, from non-relevant to relevant.

There are $2^{k_{switched}}\begin{pmatrix} N_{ranked\_doc.} \\ k_{switched} \end{pmatrix}$ different possibilities for relevance judgments within the $N_{ranked\_doc.}$ ranked document list, where $k_{switched}$ presents the number of the switched judgments.

In the first case, we assume that $k_{switched}$=1, i.e. only one document within Nranked_doc. ranked document list is switched. We assume that switch is performed on the jth position, where $1 \le j \le N_{ranked\_doc.}$. The difference between Average Precision value where jth document is relevant, and Average Precision value when jth document is not relevant is expressed such that

$$\Delta Average\,\Pr ecision\,(j) = Average\,\Pr ecision\,(x_j = 1) - Average\,\Pr ecision\,(x_j = 0) \quad (10)$$

Average Precision value in the case when jth document is not relevant is presented as

$$Average\,\Pr ecision\,(x_j = 0) = \frac{1}{N_{rel\_ret}}\left(\sum_{i=1}^{j-1} x_i \,\Pr ecision_i + \sum_{i=j+1}^{N_{ranked\_doc.}} x_i \,\Pr ecision_i\right) \quad (11)$$

On the other side, Average Precision value for the case where jth document is relevant is presented such that

$$Average\,\Pr ecision\,(x_j = 1) = \frac{1}{(N_{rel\_ret} + 1)}\left(\sum_{i=1}^{j-1} x_i \,\Pr ecision_i + \Pr ecision_j + \sum_{i=j+1}^{N_{ranked\_doc.}} x_i^{'} \,\Pr ecision_i^{'}\right) \quad (12)$$

Considering that $x_i^{'} = x_i$ for $1 \le i^{'} = i \le N_{ranked\_doc.}$ and $N_{rel\_ret} >> 1$,

$$\Delta Average\,\Pr ecision\,(j) = \frac{1}{N_{rel\_ret}}\left(\Pr ecision_j + \sum_{i=j+1}^{Nranked\_doc.} x_i^{'} \Delta \Pr ecision_i^{'}\right) = \quad (13)$$

$$= \frac{1}{N_{rel\_ret.}}\left(\frac{1}{j}\sum_{k=1}^{j} x_k + \sum_{i=j+1}^{Nranked\_doc.} \frac{1}{i} x_i^{'}\right)$$

Where $\Delta \Pr ecision_i^{'} = \frac{1}{i}$ presents the change of the Precision on the ith position where $x_i$=1. When we want to observe the change in Average Precision value when *k* switches in relevance judgment, from relevant to non-relevant are done, we can conclude that *k* simultaneous switches provoke the same changes in Average Precision value as *k* successive single switches. Therefore, the total change in Average Precision value can be presented as a total of single changes. Since that $\Delta Average\,\Pr ecision\,(j) > \Delta Average\,\Pr ecision\,(j+1)$, the value of changes is bigger when the position of the relevant document within the ranked list is lower.

## 4. Experiments

In this work, like in [14], testing was done by EBART 3. This corpus is selected as a subset of the EBART corpus 2 GB, Serbian newspaper article collection, the largest digital media corpus in Serbia. It consists only of articles from the Serbian daily newspaper "Politika", published from 2003. to 2006. There are 3366 articles in this collection. For the purpose of the first experiment, we used on small sample that were artificially generated. While, for the purposes of the second and third experiment, we use the queries collection that contains from 2KB to 6KB randomly selected newspaper articles. In order to make relevance judgment, we use "hands-on" experience in the process of evaluating information retrieval systems. Those queries we run on Google search engine. When judging pages for relevance, it should be considered that the page is relevant if it is on the appropriate topic and it is not relevant only because it has a link to a page that has relevant content. If it is not possible to fulfil those requirements, we mark the document as non-relevant. The list of 1's and 0's, presented in the Table 1 and Table 2, represents relevant (1) and non-relevant (0) documents in a ranked list of 20 documents in a response to a query. It is assumed that are 10 relevant documents in the total collection. Observing all three parameters for measuring the quality of the IR system, Average Precision, Recall and Precision, with the aim to introduce IR system performance factors sensitivity to Google ranking, in this paper, we describe three experiments.

The first experiment is artificial in order to show the IR performance dependence on the position of the relevant documents within Google ranked list. We assume that there are 20 queries and for each query document, Google's results present only one relevant document, starting from the 1st one through 20th position. This experiment is already mathematically proved in the previous section where it is explained that parameters for IR system evaluation are very sensitive to relevance judgment which is not usually very reliable. It is shown that the variation of Average Precision parameter value is highly present in the positions 1 to 10, while from the 10th position on, this variation is negligible. The results are presented in Figure 3. and it is shown that as position of the relevant documents is shifting downwards, the value of Average Precision is decreasing. Since all cases have the same scenario and the number of relevant documents within the ranked list is equal and have a value of 1, it is expected that the values of parameters for Precision and Recall are constant, while values for parameter Average Precision decline. Based on the chart, shown in Figure 3, it can be seen that the Average Precision has a more rapid decline from 1st to 10th position, while from the 10th position, this difference is negligible. The difference in values for Average Precision parameter between two successive positions is much higher at the top than at the bottom. For example, the difference in value between the 1st and the 2nd position is 0.95, while between 10th and 11th position is 0.001.

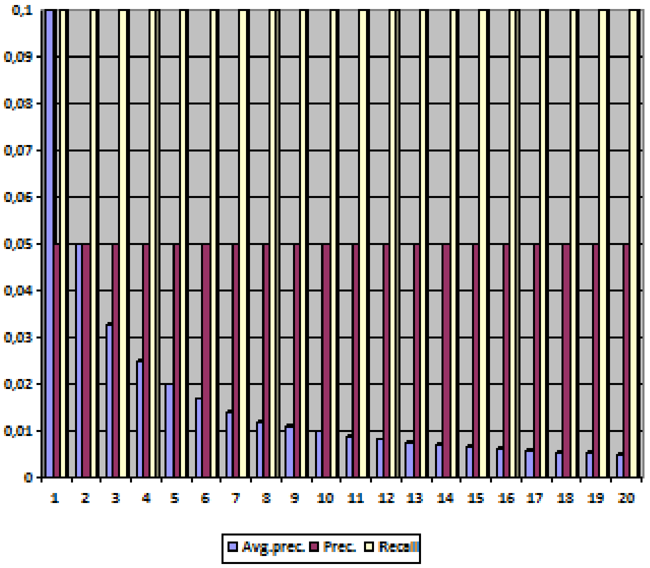

Figure 3. Dependence of the IR system performance parameters of the position of the relevant documents in Google ranked list.

Since in the previous experiment we saw that the system performance in terms of Average Precision value is better when the position of the relevant documents within the ranked list is at the beginning, the objective of the second experiment is to show that the value of the Average Precision parameter is approximately equal in those two cases: 1) only 1 relevant document in the first position of the Google ranked list, and 2) more relevant documents but at the end of this list. It is interesting to show how many relevant documents at the lowest positions are required, in order to Average Precision approximately equalize with the value of the Average Precision when the position of single relevant document is first in Google's results. First row of Table 1 shows that the first document within the ranked list is marked as relevant. The second row of Table 1 shows that the ranked list with the same number of relevant documents, with the difference that the position of the relevant document is on the last position. The values of the parameters show that the Average Precision is 20 times better in the case where the relevant document is in the first position, while the values for Recall and Precision are identical as in the case when the relevant document is in 1st position. The third row of the second table shows the relevance judgment, where the two last documents are relevant. Average Precision is 6.67 times worse than in the first case, where the only relevant document was on the top of the list, while the values of the parameters of Precision and Recall were twice as high. In each subsequent row, all values of the parameters, Average Precision, Recall and Precision are linearly increasing with increase in the number of relevant documents, while the Average Precision reaches the value of the case when the relevant document is in the first position in the ranked list only after we have 6 relevant documents at the end of the list. The chart shown in Figure 4 shows the changes in the values for each parameter.

The objective of the third experiment is to show that parameters for IR system evaluation are very sensitive to relevance judgment which can be dramatically changed during machine translation. This experiment consists of an analysis of the same query document for two different cases. In the first case, the query document is observed in original in Serbian language. In the second case, a given query document is observed as in the first case, with the difference that the query document is translated with the help of human translation first into English and then with the help of machine translation, re-translated into Serbian. Respecting the rules described above and the scenario for relevance judgment, in the first case, the result obtained shows that the relevant document in the ranked list of 20 documents is located on the 1st, 2nd, 7th, 9th and 10th position, while in the case when the query document is translated, the relevant document is on the 1st and 15th position within the Google ranked list. For obtained values of position, parameter values for the Average Precision, Precision and Recall are shown in Table 2. From the results obtained, it is evident that under the same scenario, the relevance judgment is different in the sense that in the case of translation, there are less relevant documents, and Google ranking is lower. Because of these facts, it is expected that the overall performance of IR system is worse.

Table 1. First position vs. last positions of the relevant documents

| Query | Relevance judgments | Avg. Precision | Precision | Recall |
|---|---|---|---|---|
| 1 | 10000<br>00000<br>00000<br>00000 | 0.1 | 0.05 | 0.10 |
| 2 | 00000<br>00000<br>00000<br>00001 | 0.005 | 0.05 | 0.10 |
| 3 | 00000<br>00000<br>00000<br>00011 | 0.015 | 0.10 | 0.20 |
| 4 | 00000<br>00000<br>00000<br>00111 | 0.031 | 0.15 | 0.30 |
| 5 | 00000<br>00000<br>00000<br>01111 | 0.053 | 0.20 | 0.40 |
| 6 | 00000<br>00000<br>00000<br>11111 | 0.081 | 0.25 | 0.50 |
| 7 | 00000<br>00000<br>00001<br>11111 | 0.12 | 0.30 | 0.60 |

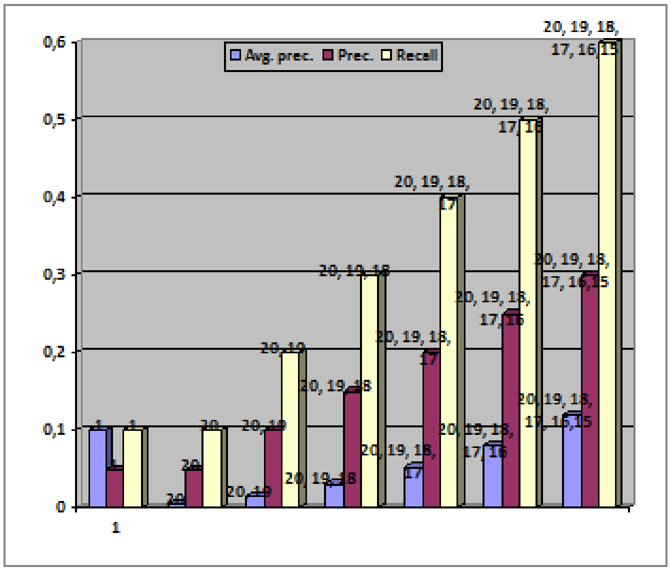

Figure 4. A comparison between the IR system performance of the parameters, in the case where the relevant document is at the beginning, and more relevant documents are at the end of the ranked list

Table 2. Experimental Results comparison for the queries for the top 20 ranked documents with and without translation

| Translated | Relevance judgments | Avg. Precision | Precision | Recall |
|---|---|---|---|---|
| No | 11000<br>01011<br>00000<br>00000 | 0.34 | 0.25 | 0.50 |
| Yes | 10000<br>00000<br>00001<br>00000 | 0.11 | 0.10 | 0.20 |

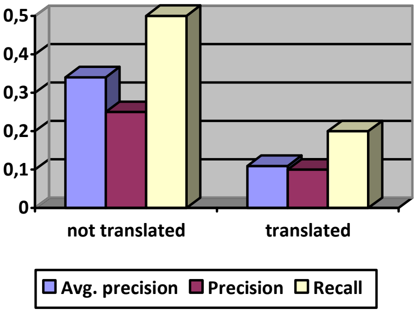

Figure 5. A comparison between the IR system performance of the parameters, in the case if the query is translated or not

### 4.1. Google Ranking

According to statistics, more than one billion people use Google search, generating about 115 billion monthly unique searches. In addition, Google occupies 67.6% market share for browsers, which means that marketers who want a high position in search results have to try hard to fulfil as much as possible Google requires. No matter how a particular document is relevant to the user query, they will be ranked according to their importance. Website is significant, for example, if there are many web pages that have a link to it.

For a typical query, there is a huge number of websites with the requested information. Today, Google Docs rely on more than 200 unique signals that allow us to guess what we are looking for [16].

## 5. Conclusions

In this paper, we observe three main parameters to measure the quality of the IR system (Average Precision, Recall and Precision). We show some interesting findings of the properties of those measures, using Mathematical and empirical analysis;

Mathematically, it is demonstrated that all those parameters are very sensitive to relevance judgment which is not usually very reliable. We show that the relevant document position shifting downwards within the ranked list is followed by Average Precision decreasing. The variation of Average Precision value is highly present in the positions 1 to 10, while from the 10th position on, this variation is negligible. Additionally, we try to estimate the regularity of the Average Precision value change, when we assume that we are switching the arbitrary number of relevance judgments, from non-relevant to relevant. We demonstrates in the case of observing the change in Average Precision value when k switches in relevance judgment, from relevant to non-relevant are done, we can conclude that k simultaneous switches cause the same changes in Average Precision value as k successive single switches. Therefore, the total change in Average Precision value can be presented as a total of single changes and change in Average Precision has bigger value when the relevance judgment switch is performed on the higher ranking positions.

Empirically, it is shown hat 6 relevant documents at the end of the 20 document list, have approximately same Average Precision value as a single relevant document at the beginning of this list, while Recall and Precision values increase linearly, regardless of the document position within the list. This way, influence of detecting unknown relevant document and its position within the ranked list on the score is discussed.

Also, we show that in case of Serbian-to-English human translation query followed by English-to-Serbian machine translation, relevance judgment is significantly changed and therefore, all the parameters for measuring the IR system performance are also subject to change.

## Acknowledgements

I would like to express my special thanks of gratitude to the Dean of Singidunum University (Professor Dr. Milovan Stanišić) who gave me the opportunity to do this project and to all my colleagues at the Singidunum University continually gave me useful advices and moral support.

## REFERENCES

[1] Languages used on internet. [Available at http://en.wikipedia.org/wiki/Languages_used_on_the_Internet]

[2] Marjanović, M., Tomašević, V. & Živković, D. (2015) "Anti-plagiarism software: usage, effectiveness and issues", *Proceedings of International Scientific Conference of IT and Business Related Research*, SINTEZA 2015, Ed., pp. 119-122.

[3] Clough, P. & Sanderson, M. (2013) "Evaluating the performance of information retrieval systems using test collections" *Information Research, 18(2) paper 582.* [Available at http://InformationR.net/ir/18-2/paper582.html]

[4] Saracevic, T. (1995) "Evaluation of evaluation in information retrieval", In Edward A. Fox, Peter Ingwersen, Raya Fidel (Eds.). *Proceedings of 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Seattle, Washington, USA.

[5] Robertson, S.E. & Hancock-Beaulieu, M. (1992) "On the evaluation of information retrieval systems information processing and management", Ed., pp. 457-466.

[6] Harman, D. (1995) "Overview of the third text retrieval conference (TREC-3)", *Proceedings of the Third Text Retrieval Conference*, Ed., pp. 1-20.

[7] Harman, D. (1995) "Overview of the fourth text retrieval conference (TREC-4)", *Proceedings of the Third Text Retrieval Conference*, Ed., pp. 1-20.

[8] Douglas, O. & Hackaett, P., (1997) "In proceedings of the sixth retrieval conference (TREC-6)", Gaithersburg, MD: National Institute of Standards and Technology, Ed., pp. 687-696.

[9] Pereira, R. C. (2010) "Cross-languages plagiarism detection", Universidade Federal do Rio Grande do Sul, Instituto de Informatica, PhD Dissertation. [Available at https://www.lume.ufrgs.br/bitstream/handle/10183/27652/000763631.pdf?sequence=1.]

[10] Milošević, N., "Stemmer for Serbian language", Natural language processing. [Available at http://arxiv.org/ftp/arxiv/papers/1209/1209.4471.pdf.]

[11] Stemmer for Serbian language. [Available at http://www.inspiratron.org/SerbianStemmer.php]

[12] Delić, V., Sečujski, M. & Kupusinac A., "Transformation-based part-of-speech tagging for Serbian language", Recent Advances in Computational Intelligence, Man-Machine Systems and Cybernetics. [Available at http://www.wseas.us/e-library/conferences/2009/tenerife/CIMMACS/CIMMACS-15.pdf]

[13] Character Encoding Recommendation for Languages. [Available at http://scratchpad.wikia.com/wiki/Character_Encoding_Recommendation_for_Languages.]

[14] Graovac, J. (2014) "Text categorization using n-gram based language independent technique", Natural Language Processing for Serbian - Resources and Applications, Proceedings of the Conference 35th Anniversary of Computational Linguistics in Serbia, ISBN: 978-86-7589-088- 1, Ed., pp. 124–135.

[15] Buckley, C. & Vorhees, E.M. (2000) "Evaluating evaluation measure stability", *In Proceedings of the 23$^{rd}$ Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Ed., pp. 33-40.

[16] Dean, B. (2015) "Google's 200 Ranking Factors: The Complete List". [Available at from http://backlinko.com/google-ranking-factors.

## Author

Marina Marjanović-Jakovljević received the Electrical Engineer degree from Belgrade University, Serbia, in 2002. In 2007. She received Ph.D. degree in Telecommunications at the Signal Processing Group of the Polytechnic University de Madrid (UPM). She has been awarded a Telefonica Moviles Fellowship to the best academic trajectory. She is currently working as an Associate Professor in the department of the Computer Engineering at the Singidunum University in Belgrade. Her research interests include Information Retreival Systems, UWB systems, ad hoc networks, and wireless communications.