On the Evaluation of Snippet Selection for Information Retrieval

A. Overwijk, D. Nguyen, C. Hauff, R.B. Trieschnigg, D. Hiemstra, F.M.G. de Jong
Twente University
arnold.overwijk@gmail.com, dong.p.ng@gmail.com, c.hauff@ewi.utwente.nl, trieschn@ewi.utwente.nl, hiemstra@cs.utwente.nl, f.m.g.dejong@ewi.utwente.nl

Abstract

In this paper we take a critical look at the evaluation method of WebCLEF 2007. The suitability of the evaluation method can be seen from two sides, namely from a participating system and a non participating system. A participant has the advantage that the evaluation is partly based upon his output. In this paper we will investigate if the size of the pool of snippets, the implementation of the evaluation method and the quality of the assessments is sufficient enough for reliable evaluation. Unfortunately we have to conclude that the evaluation is not suitable. Therefore some alternative evaluation methods will be discussed concluding in a recommendation to improve the evaluation of WebCLEF.

Categories and Subject Descriptors

*H.3 [**Information Storage and Retrieval**]: H.3.4 Systems and Software*

Keywords

Measurement, Performance, Experimentation

1. Introduction

Nowadays the information on the net is enormous and this amount is ever growing. This makes search engines a vital tool for finding information. Since almost every query gives more results than the user would be able to read, it is essential that most relevant documents are shown first. But most of the time these documents contain only a small amount of relevant information for the user. Therefore the user often has to take a look at the whole document for only a single snippet of information, which is very time-consuming. Moreover search engines generate snippets (i.e. the query terms that appear in the document and the words around them) to give the user a 'sneak preview' of the document's content [1].

Snippets help the user to assess the relevance of the document without accessing it. The generation of snippets is mainly based on the query terms and less on the context of the document (i.e. fragments in the document that do not contain query terms). Unfortunately this is only sufficient for making relevance decisions [2], which is the intention of search engines. However when the user is unfamiliar with the topic or seeks background information, context has shown to be more important [3]. This means different kinds of snippets are useful for different goals. An overview of the different search goals of the users of search engines can be found in [4].

The main reason for users to search is to 'find out about' their search topic. Therefore snippets should be more based on the context of the document. Such a goal is known as undirected information search. With over 23% of all queries, undirected information search goals are most common [4]. The WebCLEF task is about generating snippets for undirected information search. In the best case this would obviate the need to open any document, because the snippet answers to the information need of the user.

WebCLEF evaluates cross-language retrieval systems in a web setting. The task started in 2005 with navigational queries and shifted to informational queries between 2006 and 2007. The new WebCLEF task required a new evaluation method. Therefore it is good to take a critical look at the evaluation method of WebCLEF 2007 and determine if it is suitable enough to be used again in the future. A good evaluation method should be as independent as possible to the output of participating systems, which makes the evaluation reusable for non-participating systems as well. This raises the following questions:

- Is the evaluation method of WebCLEF 2007 suitable for participating systems?
- Is the evaluation method of WebCLEF 2007 suitable for non-participating systems?

We will take a critical look at the evaluation script as well at the dataset of WebCLEF 2007.

In this paper first an overview of the evaluation method of WebCLEF 2007 will be given, followed by an experimental setup to determine if the evaluation method is suitable. Then the results of these experiments will be discussed and we also take a quick view at some alternative evaluation methods, resulting in a conclusion and future work.

2. Evaluation method of WebCLEF 2007

The purpose of the WebCLEF track is supporting a user who is an expert in writing a survey article on a specific topic with a clear goal and audience. The support will consist of a ranked list with relevant snippets. The degree to which the information need is satisfied by the user is measured as the number of distinct atomic facts that the user includes in the article after analyzing top snippets returned by the system.

2.1. Data

WebCLEF provides the following information about the needs of the user:

- Short *topic title*.
- Free text *description* of the goals and the intended audience of the article.
- A list of *known sources*: online resources that the user considers to be relevant to the topic and from which information may already have been included in the article.
- Optional list of Google *retrieval queries* that can be used to locate the relevant information; the queries may use site restrictions to express user's preferences.

For every query also the top 1000 hits from Google is available. WebCLEF also provides some information about these documents:

- The *rank* in the Google result list.
- The Google *snippet*.

An example of a topic (i.e. topic 1 of WebCLEF 2007) is given below:

- topic title: `Big Bang Theory`
- description: `I have to make a presentation about Big Bang Theory for the undergraduate students. I assume that the students have some basic knowledge of physics.`
- known sources: `http://en.wikipedia.org/wiki/Big_Bang; http://www.big-bang-theory.com/; http://liftoff.msfc.nasa.gov/academy/universe/b_bang.html`
- retrieval queries: `big bang; big bang theory`

2.2. Measures

The data described in the previous section is used by participants to generate a ranked list of snippets per topic. The output of all participating systems forms a pool with snippets. Per topic the creator (i.e. the participant that created the topic) marks all the relevant parts in the pool of snippets. These relevant parts form the manual assessments, which will be used to evaluate the systems. For evaluation only the first 7000 bytes of the system is taken into account, resulting in recall and precision [5]. The *recall* is defined as the sum of character lengths of all spans in the response of the system linked to nuggets (i.e. an aspect the user includes in his article), divided by the total sum of span lengths in the responses for a topic in all submitted runs. And *precision* as the number of characters that belong to at least one span linked to a nugget, divided by the total character length of the system's response.

3. Experimental setup

In this section experiments will be described to determine if the evaluation method is suitable for participating systems as well as for non-participating systems. It is possible that the evaluation method is only suitable for participating systems, because for example the evaluation relies too much on the output of the participating systems. Non-participating systems have the disadvantage that a snippet can be marked as non-important, while it is in fact important. This happens when none of the participating systems delivered that snippet, causing it not to be included in the assessments. This does not influence the evaluation of WebCLEF 2007, but it only affects the reusability of the dataset. Several experiments will be carried out to determine whether the evaluation is suitable or not.

3.1. Pool of snippets

WebCLEF 2007 had only three participants, which can indicate that the pool of snippets was not large enough and therefore the assessments may rely too much on the participating systems. The best way to prove that this is indeed the case, would be a proof by contradiction: Assume that the dataset is reusable and show that it is not. Since the code of the best participating system of last year is made freely available, this can be proven by improving a part of the system and showing that the performance decreases. The only question left is what part of the system can be improved so that we are certain that it is definitely an improvement.

The answer to this question lies in the fact that we found a bug in the best performing system of last year. This bug affects the filtering of stop words, which is a part of the *similarity-based centrality* ranking algorithm of the system [6]. Due to a map implementation instead of a list, only half of the stop words are filtered at random. This is because a map consists of key-value pairs, where the odd words are the keys with the even words as their values. In consequence only the keys are filtered, leading to filtering only half of the words. Adding for example a stop word to the beginning of the list, makes all odd words even and all even words odd and therefore the other half of the stop words will be filtered. Resulting in filtering half of the stop words at random.

We will run three evaluations, one with the bug, one without the bug and one without filtering stop words at all. This last run is necessary to determine if filtering stop words is a good approach at all. No matter the fact that filtering stop words is a good approach or not, we can be sure that one of both runs must be better than the run with the bug.

3.2. Implementation

The evaluation of WebCLEF 2007 is based upon precision and recall. With this experiment we will test if the measurement of precision and recall is correctly implemented. A system that delivers output that is almost similar to the output of a participating system should almost have the same performance (i.e. precision and recall). An experiment for testing this would be taking the output of last year's best performing system and remove the last word of every snippet. The performance of this might be somewhat lower, but removing the last word may not have a huge influence since the average length of a snippet is over 40 words.

3.3. Quality of assessments

The evaluation of the WebCLEF task highly dependents on the manual assessments. Therefore the quality of these assessments is very important. In this experiment the quality of the assessments is explored.

The ranking algorithm of the best performing system of last year is dependent on known sources provided by the user. For some topics there are no known sources available, which means that no score can be calculated. For these topics the system delivers the snippets in order of occurrence (i.e. the first snippet is the first paragraph of the first document, the second snippet the second paragraph, etc.). This would be the same as delivering the first documents, like a search engine does.

An experimental system that outputs the snippets simply by occurrence will be evaluated. The performance of this experimental system will give us an indication about the quality of the assessments.

4. Results

Results of the experiments described in section 3 are given below:

Table 1. Pool of snippets

System	Precision	Recall
With bug	0.2018	0.2561
Without bug	0.1328	0.1685
Not filtering stop words	0.1087	0.1380

We can see that the performance of the system without the bug is significantly decreased, which would indicate that filtering stop words is not a good approach, but not filtering stop words at all gives an even lower performance.

Table 2. Implementation

System	Precision	Recall
Original	0.2018	0.2561
Last word removed	0.0597	0.0758

Removing the last word in the output of the original system has a huge influence on performance, which should not be the case, since the snippets are almost similar.

Table 3. Quality of assessments

	Original		First occurrence	
Topic	Precision	Recall	Precision	Recall
17	0.0389	0.0436	0.0389	0.0436
18	0.1590	0.6190	0.1590	0.6190
21	0.4083	0.6513	0.4083	0.6513
23	0.1140	0.1057	0.1140	0.1057
25	0.4240	0.4041	0.4240	0.4041
26	0.0780	0.1405	0.0780	0.1405
Avg.	0.2018	0.2561	0.0536	0.0680

Simply returning the snippets by first occurrence gives for six out of thirty topics (20%) the same performance of last years best performing system. Remarkable is the small difference in performance between the experimental system that removes the last word and the experimental system that returns the snippets by first occurrence.

5. Discussion

According to the results of the experiments, it is clear that the evaluation method of WebCLEF 2007 is not suitable. The *pool of snippets* experiment indicated that there were not enough participating systems to create a pool that is large enough to make evaluation possible for non-participating systems as well. However the problem with the evaluation method cannot simply be explained by a too small pool of snippets, caused by the low number of participants, nor can it be explained by the quality of the assessments. Unfortunately there are several problems that caused these remarkable results.

First of all the precision and recall measurement is not implemented in the right way. After analyzing the evaluation method, it turned out that a snippet in the manual assessments should be part of the output of the system. Snippets delivered by the system that, for example, differ in only one character with snippets in the assessments do not have a match. This means that the precision and recall value will be zero. Although the snippets are almost the same and therefore contain almost the same information as the snippets in the assessments, which should lead to a high precision and recall value. The consequences of this problem are illustrated by the *implementation* experiment. Solving this problem is not an easy job, since the same information can be represented in several ways. The TREC QA task also has to deal with this problem [7].

In addition there are also some problems with the assessments. This is indicated by the *quality of assessments* experiment. For example some topics do not contain relevant snippets (e.g. topic 14) and other topics sometimes do not even contain any snippets at all (e.g. topic 12), which automatically results in a precision and recall value of zero. The combination of all these problems results in an inappropriate evaluation. The *quality of assessments* experiment shows that a system that simply returns the snippets by first occurrence performs almost equal to the *remove last word* system that produces almost similar output as last year's best performing system.

6. Alternative evaluation methods

The problem with the evaluation method is very complex and a solution is not directly available, although there has been done a lot of research to this problem already. When designing a new evaluation method, many choices have to be made. Whatever these choices will be, the method must be offering system developers much insight into its

parameters. This is already difficult enough, but in addition the performance of these systems is far from high. The combination of these facts makes it hard to develop task-oriented evaluations that are both related to the researchers' interests and are not too far beyond their systems' capabilities [8].
In the WebCLEF task, text quality (i.e. can the system produce 'proper' sentences and 'properly connected' discourse) is not that important, since only extractive approaches are used so far. However concept capturing (i.e. does the summary capture the key concepts of the sources) is more important. Unfortunately it is much harder to measure the second issue, because not only it involves judgements about conceptual importance in the source but because concepts, especially complex relational ones, are not clear cut and they may be variably expressed [8].
One of the problems, mentioned earlier, is that the assessments do not contain enough relevant snippets and therefore are not reusable. This is caused by the fact that the assessments are based upon a pool of snippets that is not large enough due to the small number of participating systems. [9] showed that for this reason measures based on precision and recall are highly uncertain. There has been done some research to cope with incomplete assessments (e.g. bpref [10]). This can be a possible solution for the reusability problem. TRECEVAL for example is a program that also reports this bpref value. Unfortunately TRECEVAL cannot be used for the WebCLEF evaluation, since it is impossible to give snippets an id. This in turn, is caused by the fact that the participating systems have to extract the snippets from the document collection their self and therefore not every system extracts the same snippets.
This leaves us with the problem of how to evaluate in general. An approach that is close to the current evaluation method, and therefore a reasonable solution, is providing the offsets (i.e. the start and end of a passage in the document) of the delivered snippets. With this information the amount of overlap can be calculated to get an indication of the performance. A similar approach is already used in XML Retrieval [11].
Another more common approach for evaluating extractive summaries, which is the case in WebCLEF, is automatic comparison between reference and system summaries using n-grams. Originally this approach is applied to machine translation, but it has been developed in the ROUGE program for summary evaluation as well [12].

7. Conclusion & Future work

First of all we can conclude that it was very valuable to take a critical look at the evaluation method used in WebCLEF 2007. For developers it is very important to measure the performance of their system. Especially in a task where it is hard to measure the quality of the output (i.e. WebCLEF). Moreover we think that it is worthy to have a critical look at evaluation methods in general, especially when the method is not already commonly used. However even when the evaluation method is commonly used, it does not guarantee that it is correctly implemented.
In this paper we have shown that the evaluation method as well the dataset of WebCLEF 2007 does not provide information that is of the researchers' interest nor does it reflect the performance of the system in a correct way. We have shown that the manual assessments were not carefully created. Problems that occur with manual evaluation is that it most of the times is very hard to judge whether a snippet is relevant to the user. Therefore some other tracks (e.g. [7]) have multiple assessors per topic, which certainly improves the quality of the assessments.
Moreover we have shown that the measurement in general is not appropriate. With the current evaluation method a snippet in the assessments must occur exactly in the system's output. This is not realistic, since the same information can be variably expressed. Unfortunately this problem also occurs in other tracks (e.g. [7]). Using n-grams (e.g. ROUGE [12]) will partly solve this problem, because it is unlikely to happen that the same information has no words in common. It might even be better to combine such an approach with the importance (e.g. frequency in the top documents retrieved by Google) of the n-gram. We leave this question for future work.
The conclusion we can draw from these observations is that the evaluation is certainly not reusable for non-participating systems, since all mentioned problems affect the evaluation for non-participating systems. For participating systems on the other hand, we have to conclude that the current evaluation is not optimal, but might be sufficient. This depends on the fact that the output of all systems are carefully assessed. The assessments for example contain for some topics multiple times the same snippet with a small difference (e.g. less words), which indicates that this snippet occurred in the output of different systems. If this is indeed the case, then it bypasses the problem with the implementation. Also the pool of snippets does not affect the evaluation for participating systems, since their snippets make part of this pool. Only the problem with the quality of the assessments cannot be bypassed, however when the assessors did their job objective, which we will assume, then it should affect all systems in the same manner.

Acknowledgements

This paper is based on research partly funded by IST project MESH (http://www.mesh-ip.eu) and by bsik program MultimediaN (http://www.multimedian.nl).

References

[1] A. Turpin, Y. Tsegay, D. Hawking, and H. Williams, E., "Fast generation of result snippets in web search," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval* Amsterdam, The Netherlands: ACM, 2007, pp. 127-134.

[2] D. M. McDonald and H. Chen, "Summary in context: Searching versus browsing," *ACM Trans. Inf. Syst.*, vol. 24, pp. 111-141, 2006.

[3] E. Carmel, S. Crawford, and H. Chen, "Browsing in Hypertext: A Cognitive Study," *IEEE Trans. Syst. Man Cybernet.*, vol. 22, pp. 865-884, 1992.

[4] D. E. Rose and D. Levinson, "Understanding user goals in web search," in *Proceedings of the 13th international conference on World Wide Web* New York, NY, USA: ACM, 2004.

[5] V. Jijkoun and M. de Rijke, "Overview of WebCLEF 2007," in *CLEF 2007*, Budapest, Hungary, 2007.

[6] V. Jijkoun and M. de Rijke, "The University of Amsterdam at WebCLEF 2007: Using Centrality to Rank Web Snippets," in *CLEF 2007*, Budapest, Hungary, 2007.

[7] E. M. Voorhees and D. M. Tice, "The TREC-8 question answering track evaluation," in *Text Retrieval Conference TREC-8*, pp. 83-105, 1999

[8] K. Sparck Jones, "Automatic summarising: The state of the art," *Inf. Process. Manage.*, vol. 43, pp. 1449-1481, 2007.

[9] J. Zobel, "How reliable are the results of large-scale information retrieval experiments?," in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval* Melbourne, Australia: ACM, 1998.

[10] C. Buckley and E. M. Voorhees, "Retrieval evaluation with incomplete information," in *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval* Sheffield, United Kingdom: ACM, 2004.

[11] J. Pehcevski and J. A. Thom, "HiXEval: Highlighting XML Retrieval Evaluation," *Advances in XML Information Retrieval and Evaluation: Fourth Workshop of the INitiative for Evaluation of XML Retrieval (INEX 2005)*, 2006.

[12] C.-Y. Lin, "ROUGE: a Package for Automatic Evaluation of Summaries," in *Proceedings of Workshop on Text Summarization*, Barcelona, Spain, 2004.