

# Evaluation Design of Information Retrieval System with eTVSM Specific Extensions

Artem S. Polyvyanyy

Hasso-Plattner-Institute for IT Systems Engineering  
at the University of Potsdam  
D-14482 Potsdam, Germany

[Artem.Polyvyanyy@student.hpi.uni-potsdam.de](mailto:Artem.Polyvyanyy@student.hpi.uni-potsdam.de)

## ABSTRACT

Enhanced Topic-based Vector Space Model (eTVSM) [21] is an advanced information retrieval model that integrates stemming and stopword removal and can represent most of the linguistic phenomena. Being a promising one, this model still lacks quantitative evaluations and comparisons to other models. This paper aims to design a plan for evaluation of a eTVSM information retrieval system. The effectiveness of information retrieval systems is measured by comparing performance on a common set of queries and documents. Significance tests are then used to evaluate the reliability of such comparisons. This paper describes the preparation of the test collections together with following experimental procedures, measurements explanation and statistical evaluation of the results. Thus, gives the complete instructions set for performing evaluation of information retrieval model and in particular eTVSM.

## Categories and Subject Descriptors

H.3.4 [Information Storage and Retrieval]: Systems and Software --- *Performance evaluation (efficiency and effectiveness)*

## General Terms

Experimentation, Measurement, Performance

## Keywords

Significance test, quantitative evaluation, information retrieval, Enhanced Topic-based Vector Space Model

## 1. INTRODUCTION

Information retrieval is the method of searching information in documents, documents themselves or metadata that describes these documents [1]. This definition is not dependant on method of document storage, or their type which determines the content of information being searched. This can be a search in the local database or in the Internet for text, images, sound, or data.

Information retrieval is a loosely-defined term and the problem of

information retrieval can be investigated under different aspects. This paper deals with the automatic information retrieval tasks of the information represented as text. Moreover, we will deal with the method of comparison of different information retrieval models. In addition, we will discuss the approach to evaluation of eTVSM.

There are a number of described and deployed information retrieval models. Some of them are quite simple and do not require much computation power, others are much more packed with algorithms claiming to deliver better results. In addition, new and promising models appear on the market. In this paper we discuss the method of performing quantitative evaluation of information retrieval models. As a consequence we can come up with an order for them through pair-wise comparison. In this case we can say which system performs better in the given conditions.

The picture becomes a little more complex if, instead of two alternative systems, we have one system which we think might be capable of improvement. In this situation, we might for instance want to evaluate how well it performs under specific model setup and how it reacts to the changes of the setup variables. This is especially important when evaluating a model which performance highly depends on its internal configuration. This is exactly the case with eTVSM. eTVSM claims to be able to represent most of the linguistic phenomena. However, how well it copes with this task highly depends on the configuration of the topic map eTVSM computes upon. Nevertheless the process of model configuration is highly heuristic. In this paper we try to come up with a step by step information retrieval model evaluation technique that should give the better understanding of the system effectiveness as a response to its configuration.

In this paper we propose a complete methodology for the information retrieval system evaluation. It includes the complete description of all steps that should be taken to perform quantitative evaluation of the information retrieval model. It includes preparation of the workload, following by actual experiment description with further statistical aggregation of the collected measurements.

In general, the proposed methodology is suitable for the evaluation of any information retrieval model. The described statistical analysis is not only suitable for the comparison of two different information retrieval models, but can also be applied for the study of the model effectiveness under different model configurations.

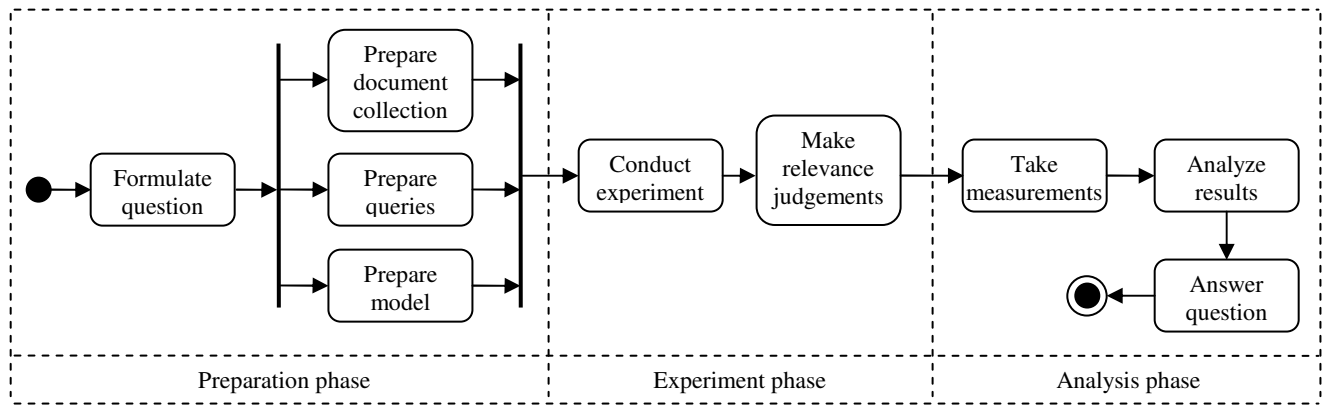


Figure 1. Evaluation design.

We start this paper by defining the evaluation type (section 2) we are going to describe. Further, in section 3 we specify evaluation design, which are the steps and their order that should be carried out to perform evaluation of the information retrieval system. In sections 4 and 5 we describe the preparation phase of the proposed evaluation design. Section 4 deals with the preparation of the test collection, while section 5 is dedicated to the setup specific for information retrieval model used – in particular eTVSM. Afterwards, in section 6 we present experimental design following by the description of the measurements one might be interested in while performing evaluation in section 7. Finally, in section 8 we present two approaches to statistical aggregation of observed measurements with further study of the significance of the received results. We sum up this work with the conclusions. We also give a hint on a future work that should follow the work presented in this paper.

## 2. EVALUATION TYPE

### 2.1 Operational vs. Laboratory

Operational and laboratory type of tests contradict each other [1]. On the one hand one can perform a pure laboratory test keeping all the system variables under control. By doing this, we can exclude the possibility of any extraneous variations that may influence the results. This can be seen as ideal test conditions. On the other hand one might be interested in how these results are applicable to the real world problems of the design of information retrieval systems. Also, to determine whether these results provide answers which will be applicable in real situation a test must be conducted in an operational environment.

This conflict results a wide spectrum of testing methods ranging from pure laboratory experiments to a study of real systems and users in real operating environment. We are not going to solve this conflict, but we will perform a laboratory test. To motivate our selection we will look for the analogy in social science. While performing questionnaires sociologists look for the representative selection of participants out of the complete set and then further generalize on the results received. Failing to come up with an appropriate selection will discard the output. A bright example of such misunderstanding was performing at the beginning of the 20<sup>th</sup> century a poll in the United States about the preferences of the citizens prior to a presidential elections. The poll was conducted over the telephone. Failing to realize that only rich people used telephones at that time led to the wrong predictions.

Unfortunately the chance to realize this bias happened only after the elections.

Therefore, we will perform a pure laboratory test under the assumption that our test setup is configured to correspond to the real world model and therefore statistical generalization is adequate.

### 2.2 User Centered vs. System Centered

To continue the idea of study of the real user's information retrieval system operating environment evaluation we can distinguish between user centered and system centered evaluation. Unlike the previous classification these two approaches are not in conflict, moreover they supplement each other.

User centered evaluation approach deals with the usage convenience issues of the information retrieval system. This is independent of the model that is being used. Such conveniences might include intuitive user interface, acceptable computation delays, and full support of the user workflow with the system.

On the other hand, system centered evaluation deals with the computational aspects of the model. The questions one should be concerned with here are how accurate are the results and how cheaply can the algorithm deliver them.

In this paper we describe system centered approach as we are interested particularly in the algorithmic "quality" of the model. However, the importance of user centered evaluation should not be underestimated as perfect content can be neglected by poor presentation.

## 3. EVALUATION DESIGN

In this section we will discuss the design of the overall evaluation setup. In other words what steps and in which order should be taken in order to perform evaluation. In the Figure 1 you can see the proposed big picture of evaluation design that is given in the notion of activity diagram.

Tests in general, and experiments in particular, are normally intended to answer specific questions. In our case the question that we want to answer is: "How effective is the information retrieval system?"

*Effectiveness is how well the system does what it is supposed to do; its benefits are the gains deriving from what the system does; its efficiency is how cheaply it does what it does [1].*

In this paper we will understand effectiveness as the ability of the information retrieval system to retrieve relevant documents and suppress non-relevant documents. In this context we can assume benefit to be proportional to the effectiveness. The efficiency of the system is out of scope of this paper as it can be evaluated as a part of the user centered approach.

Therefore, as you can see in Figure 1, we start with the question and then perform steps in order to be able to answer this question at the end.

### 3.1 Elements of the Evaluation

We would like to come up with approach to perform evaluation of the information retrieval system. Moreover, we would like to make this approach suitable for evaluation of eTVSM and make it such that will aid us to choose optimal strategy for eTVSM configuration. Further we will propose short description of the basic phases that we propose for the evaluation process.

The preparation phase deals with getting ready all the components needed for performing evaluation. The actual system evaluation will start in the next phase; however, it is preferable that prior all necessary measures are taken. These measures include getting the workload for the system that will be tested as well as preparation of the system itself. All the internal variables that influence system work should be configured to the combination that is a target for performing evaluation. In general, we should provide such evaluation setup at this phase that will aid us to answer question which we are interested in.

An important component of any test is the experimental design - the way in which the test is organized in order to answer our questions that we pose to the system. At this stage all the information prepared in the previous phase should be effectively fed into the system. By saying effectively, we mean the way that allows completely exploiting all available information, and transforming it into the data array that is suitable for statistical analysis. We do this by collecting measurements in the course of the evaluation for every single experiment that accumulates the knowledge about this experiment and that characterizes the system in the way it is needed for the evaluation.

In the analysis stage we use statistical methods to work with measurements collected during experiment phase. This is a pure statistical phase, where by statistical means we try to understand what does the retrieved results say about the system, and to which extent they can be generalized. After this stage either we answer the question, or decide that the data processed was not informative enough to aid us in answering the question and that the further steps should be taken to finalize evaluation.

## 4. TEST COLLECTION

To conduct a performance evaluation test of the system we need to come up with the appropriate workload. In case of information retrieval system this workload is represented in the form of a test collection. Further, we will discuss the content of such test collection.

## 4.1 Overview

How does a regular interaction of a user with the information retrieval system look like? The user wants to retrieve data from a large array of documents. To do this he/she formulates information demands in the form of a query/request. Afterwards, the user feeds this query into the system and receives relevant documents retrieved. The relevance here is considered by the information retrieval system. From this simplistic example it is clear that we need a document collection upon which our system will perform retrieval and queries that will initiate retrieval process. This already enables us to make the system work – perform some actions and return results. However, these results are pointless without actual users deciding whether documents retrieved by our system are really what he was looking for. To be able to judge the work of our system we need to have relevance judgments or relevance assessments that can be seen as the following function:

$$rel(d, q) \in [0..1]$$

where  $d$  – is a document from our document collection and  $q$  is a query. In general this function is defined on the interval from zero to one. “0” - means that this document is absolutely not related to the query, and “1” meaning that this document is absolutely related to it. Other values in the interval reflect upon the relevance level. As a special case we can work with the  $rel$  function that is defined in the set  $\{0,1\}$ .

At the end these three components: document collection, queries and relevance judgments form the test collection. We will now describe each of these components.

### 4.1.1 Document Collection

There are plenty of documents available on the web: news articles, electronic libraries, web pages. All of them suit for our purpose.

Are there alternatives? It is possible to do some simulation experiments using pseudo-documents which are generated in some fashion (perhaps involving Monte Carlo techniques) [1]. This kind of simulation of course has its role and can aim answering specific research questions. But, in most cases it is easier and better to use genuine documents. There is of course no shortage of those. Or, one might use an existing collection of documents which is used for information retrieval needs.

### 4.1.2 Queries

The situation with queries is much more problematic then in case with documents. And there are number of reasons for this. If we stick to the idea of simulating real world condition we need to obtain real queries. The problem with genuine queries is that it is hard to trap them as they exist for a short period of time and the location of such request acts is usually sparse. But this is nothing in comparison with the actual time needed to collect the queries that are representation of anything.

In case of the document collection which is a target of information retrieval activities this approach is still acceptable. It is not clear what to do in case of arbitrary approach document collection creation when potential users have no idea of documents being included in the collection. Can we guarantee that genuine queries will at least target the topics of the documents in such collection? Here arises a strong necessity in artificial queries construction. Such artificial queries may vary in their degree of realism; the

main point is that they should exploit the document collection properties. The problem here is that we do not really know what are the important characteristics which we should be trying to reproduce [1].

It is critical to understand the importance of queries selection. These are queries that will initiate retrieval process of the evaluated system. Thus, the obtained queries collection should be able to numerously exploit all the aspects of the retrieval model to allow statistical generalization on the results. In order to achieve this goal queries should be formed to be matched to the most of the documents from the document collection and carry different levels of relevance to these documents.

### 4.1.3 Relevance Judgments

The definition of the effectiveness as we have defined it is good for the understanding of what is happening but is definitely not sufficient as a formal basis for an experiment.

How can we come up with the appropriate relevance judgments? As the output of the information retrieval system is targeted for the human use it is only the user who can make such judgments. The next question which arises is which user's judgments can be considered correct? Usually people have different preferences and different point of views on the common issues.

Therefore, the relevance judgments cannot be done by a single person but should be performed by a group of people simultaneously on the same query on the same document collection. Further the estimated measured parameter should be obtained as the mean value over calculated measurements from all the judges. First of all this approach allows us to construct confidence interval for the measurement estimator and make further reasoning on the number of the judges being sufficient. Also, we can classify queries that cause judges to provide sparse relevance assessments. Finally, this measurement obtaining approach allows us to introduce valuable assumptions about measurements arbitrary nature which we will discuss later.

Relevance judgments can be considered as a bottle-neck of the overall test collection creation process. Ideally, the relevance judgments should be made for each document upon each query. Considering the fact that each query-document pair is judged by several people the amount of work needed to be done explodes. Further, by assuming an ideal case when judgments are made on the full content of the document this work can be out of manageable scope for the experimenter. Due to this reason we advise to take a look for existing test collections that were used for other information retrieval experiments and are publicly available online.

## 4.2 Existing Test Collections

By now, it should be clear that the test collection creation is itself a challenging task. Additionally, failing to come up with representative test collection sample may, and probably will spoil all further calculations. The worst part here is, that you, as an experimenter will have no clue about this. Therefore, it makes sense to first take a look for existing test collections that have been used in different information retrieval experiments and are publicly available. These collections are already tested and are proven to deliver usable results.

In general, we would like to classify existing test collections into three categories: classical, comparatively new publicly available and paid.

In classical test collections we include those that were used in the first experiments of information retrieval and still appear in the literature as those that were useful in the past.

In Table 1 you can find a list of such most commonly known test collections

**Table 1. Classical test collections**

Name	Docs	Qrys	Size
ADI	82	35	0,04
Time	423	83	1,5
Medline	1033	30	1,1
Cranfield	1400	225	1,6
CISI	1460	112	2,2
CACM	3204	64	2,2
LISA	5872	35	3,4
NPL	11429	93	3,1

*Docs – number of documents, Qrys – number of queries, Size – size in megabytes.*

All of the test collections provided in Table 1 [26] are available online and consist of the components described above in this paper. Here we should mention that ADI test collection is rather small and therefore any statistical analysis on it will not bring significant results.

As the next category we distinguish test collections that appeared later. These collections are usually larger in size. The first one that can be classified to this category is Reuters-21578. It consists of 21578 documents and is the ancestor of Reuters-22173 [9]. These collections are approximately 20Mb in size. Long time Reuters-21578 was most widely used in text categorization evaluations, but now the situation seems to change in favor of larger collections such as Reuters Corpora Vol.1 and Reuters Corpora Vol.2. Though all these test collections were originally designed for information filtering evaluations and thus have no queries provided for conducting information retrieval evaluations, they are widely used for retrieval evaluations as well. The approach for creating queries is described in [9]. Also, during TREC 2002 [3], 50 search topics (test queries) were developed on Reuters Corpora Vol.1 for the filtering track. They can be applied for testing ad-hoc search [10].

Reuters-21578 is available for download on the web [25], while Reuters Corpora Volumes one and two are available from NIST, the National Institute of Science and Technology. Another serious of test collections is available from National Institute of Informatics [11].

Finally, we want to draw your attention to the test collections provided by TREC [3]. The Web research collections are distributed by the University of Glasgow for research purposes only. In order to receive copies of one or more of these

collections, you must sign an agreement with the University of Glasgow and pay a contribution to the University's various costs in preparing and distributing the data. These collections and contribution amounts are provided in Table 2.

**Table 2. Paid test collections**

Name	Size	Fee
WT2g	2 Gb	£250
WT10g	10 Gb	£400
.GOV	18 Gb	£400
.GOV2	426 Gb	£600

### 4.3 Suitable Test Collections

We recommend that you use any of the provided test collections. The priority should be given in the order the test collections appear in the paper. With the small test collections experiments can be held not just on one but on several collections.

For the collections that do not completely satisfy test collection composition, as in case of Reuter's collections, additional components should be provided prior an experiment.

The work of Blair and Maron [12] has indicated that retrieval performance varies with collection size. Their findings suggest that results from experiments based on small test collections may not hold when applied to larger collections. Due to this classical test collections are now rarely used with information retrieval researchers.

In case you want to create your own test collection you need to consider all the issues mentioned in this section to make it representative and therefore usable for information retrieval experiment.

## 5. INFORMATION RETRIEVAL MODEL SETUP

At this stage we should have prepared the workload for the system evaluation. Prior we can start with experiments the system itself should be prepared.

The preparation of the information retrieval system should only deal with the system's internal variables that influence the retrieval process of the system. For some information retrieval models such as "pure" Vector Space Model (VSM) there is no need in this step at all. The model output depends only on the system input and does not depend on the model configuration. From this point of view, such model can be called stateless, it encapsulates algorithm that operates only on the input.

Unfortunately, this is not true for all the information retrieval models. Many models use additional access to some form of the knowledge databases to improve retrieval quality. One of such examples is eTVSM which uses ontology. Different ontology configurations will deliver different output in the form of retrieved documents. Therefore, it becomes interesting which ontology configuration tends to deliver better results. Of course, this question can be formulated not only in respect to ontology but

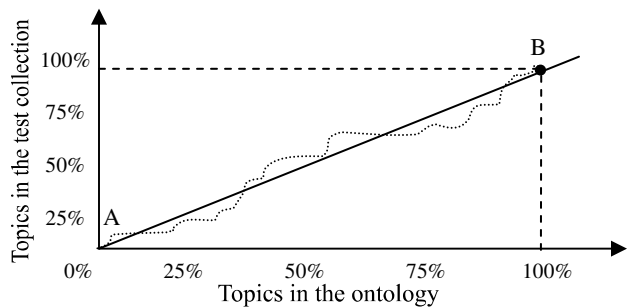
to other configurable layers of eTVSM as well. In this paper we discuss an approach to the study of ontology configuration influence on the efficiency of eTVSM.

### 5.1 eTVSM Setup

In general, the approach proposed in this paper is suitable for any information retrieval model. However, information retrieval model setup step is of course model specific. In this paper we discuss the approach to eTVSM evaluation, for this reason in this section we will concentrate our attention on the suitable configuration strategy for performing evaluation in order to be able to find the optimal eTVSM setup.

For a description of eTVSM please refer to [21]. In this paper we assume your basic knowledge of this model. The document similarity calculations depend on interpretation, topics and topic map layer. It is assumed that documents are mapped on the interpretation layer with the general rules of language particularities. Further, interpretation and topics layers should be configured in the way to achieve major benefits of eTVSM – representation of the most linguistic phenomenon. Though it is possible to study the effectiveness of the eTVSM under the different configurations of the above mentioned layers this is out of scope of this paper. Thus, the influence of the topic map (ontology) configuration is of interest.

The proposed approach to study of the influence of ontology configuration on the effectiveness of eTVSM is very simple. We start with the empty ontology. Then, step by step we add topics to the ontology until we cover all the topics from the test collection. Further, we do not stop but continue with refining topics in the topic map. Visualization of this process is proposed in Figure 2.



**Figure 2. Ontology development.**

Point A is the starting point when ontology is empty. Point B represents the topic map that includes all topics from the test collection. Points on the line that connects A and B represent intermediate configurations which include subsets of topics from the test collection. The solid line represents the shortest path from A to B while dotted line is one of the possible realizations of the topic map step by step refinement.

Now, the question is where to take topics from? We need to have a set of all topics represented in the test collection. We propose to extract these topics from queries that are the part of the test collection. Documents in the test collection are the representation of the target real world documents topic distribution. By including queries in the test collection we specify the experiment – how the system should be evaluated. Queries should exploit document

collection in full, thus should include the topics of the document collection. Therefore, system will be mostly evaluated for the topics included in queries.

The simplest way to extract topics from queries is to take all the meaningful words out of them. Further, we can start with the step by step topic map refinement. There are several approaches possible:

- Starting from the most general topics with further refinements to the most specific ones;
- Starting from the most specific topics, afterwards generalizing on them with more broad topics;
- Starting by representing some queries in full while others are not included at all.
- Adding topics by simultaneously using topics from all queries.

The actual strategy can be any combination of the proposed ones. Moreover, the strategy can vary at different points in the evaluation process. However, we advise to use the strategy of the most common topics refinement by simultaneous usage of topics from all queries. By introducing this approach you simulate natural ontology development which is test collection specific (excluding the concepts that are not the target of the evaluation).

Another topic map modeling issue is the topic relations used to construct topic map: is-a, consists-of, other or a mixture of the relationships. Anyway, this can be seen as a target for eTVSM evaluation. Evaluation can be performed for a pure relationship usage or any combined strategy using the same set of topics available for modeling.

If with the starting point of the topic map everything is clear, the end point needs additional discussion. As the topic map develops the evaluated parameter asymptotically approaches its true value. This value is the limit value that the evaluated system can produce. The visualization is provided in Figure 3.

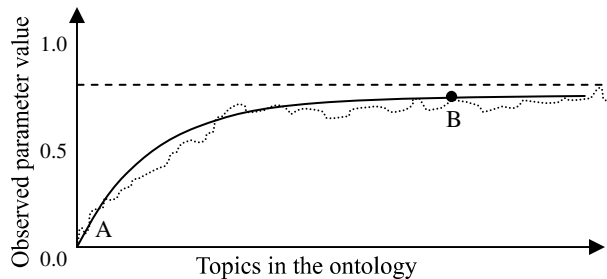


Figure 3. Observed parameter evolution.

Here, the solid line represents the theoretical value of the observed parameter. The dashed line is the limit of the parameter for the evaluated system. The dotted line is one possible development of the evaluation. The end point in this sense is the one after which further topic map refinements do not produce much gain in the improvement of observed parameter. Thus, the effort needed to reach that value is too high in comparison to the benefit of the system effectiveness gain. This point should be considered as the end point.

The evolution of the topic map starting from point A up to point B was already discussed. The end point of the evaluation might happen at any moment. Theoretically, this moment might happen before point B is reached. Just for the same reason, there might be the need to continue evaluation after point B. In this case, topics already represented in the topic map should be further refined. This can be done by extracting topics from document collection or simply by generalizing or refining topics from the topic domain of the test collection. Anyway, this process cannot be unambiguous and is highly heuristic.

Therefore, actual evaluation deals with a step by step topic map refinement with further evaluation of observed measurements. By collecting configuration-measurement pairs we are able to come up with the graphic that looks similar to the one shown in Figure 3. Further, the level of appropriate topic map saturation one might conclude by finding common tendencies over observed measurements or over test collections.

## 6. EXPERIMENTAL DESIGN

To support the workflow proposed in the beginning of the paper, which is shown in Figure 1, experimental design should bind preparation phase with available workload, configured system and acquired measurements.

Even having such a narrow scope, it has many aspects. And it is not possible to generalize on the experimental design for information retrieval systems evaluation in the wide sense. Most of such existing experiments use just one set of requests/queries to evaluate or compare a number of systems. It is called 'matched pairs' procedure [1], when the efficiency of the systems is compared on the same request. Moreover, there is a clear statistical reason for such approach. Any statistical significance testing will be much more efficient with this method. Again, this approach is oriented to decrease the influence of the bottle-neck of the whole evaluation process which is the amount of requests. With this approach it is possible to reuse the requests decreasing the need in the larger number of distinct requests.

Therefore, experimental design can be quite simple. Each request/query is searched against every system or every system configuration. Since the searching part of the system is controlled by simple rules, there is no problem in relation to replicating searches or the order in which the systems are tried [1]. The only matter of convenience in case of a single system evaluation is performing the evaluation for the whole request set for a single configuration, further reconfiguring the system and performing the complete course of evaluation for a new setup.

## 7. MEASUREMENTS

Now, we know how to perform experiment. In order to be able to answer the question we have posed at the beginning we need to perform statistical evaluation of the measurements taken in the course of experiment. There are number of ways to measure how well the retrieved information matches the intended one. We will use the standard recall, precision and F measures.

Let  $D$  be a set of documents for validation and  $K$  be a set of criteria/queries.

Let  $R_k^S \subseteq D$  be a set of documents relevant to the system, and  $CR_k^S = D \setminus R_k^S$  be a set of documents, not relevant to the system.

Then, analogously we can define  $R_k^T \subseteq D$  as a set of documents, relevant according to the user feedback and  $CR_k^T = D \setminus R_k^T$  a set of documents not relevant according to the user feedback.

In these terms recall can be defined as the ratio of correct assignments by the system divided by the total number of correct assignments:

$$R_k = \frac{\#(R_k^S \cap R_k^T)}{\#R_k^T} \in [0..1]$$

in other words,

$$R = \frac{\text{number of relevant documents retrieved}}{\text{number of relevant documents}}$$

Recall measures the completeness of the results; therefore, high values are desirable.

Precision is the ration of correct assignments by the system divided by the total number of the system's assignments:

$$P_k = \frac{\#(R_k^S \cap R_k^T)}{\#R_k^S} \in [0..1]$$

$$P = \frac{\text{number of relevant documents retrieved}}{\text{number of documents retrieved}}$$

Precision measures the accuracy of the results; therefore, high values are desirable.

Recall and Precision can also be evaluated at a given cut-off rank, denoted as  $R@n$  and  $P@n$ , instead of all retrieved documents.

Another standard measure is the F-measure, which is the weighted harmonic mean of precision and recall. F measure can be obtained as:

$$F = \frac{2pr}{(p+r)}$$

It combines recall ( $r$ ) and precision ( $p$ ). This is also known as  $F_1$  measure because recall and precision are evenly weighted.  $F_1$  measure was initially introduced by van Rijsbergen [13].

In general case the formula is given for arbitrary  $N$  as:

$$F_N = \frac{(1+N^2)pr}{p+N^2r}$$

Two other commonly used  $F$  measures are  $F_{0.5}$ , which measures precision twice as much as recall, and  $F_2$  measure, which weights recall twice as much as precision.

As the negative measure of the system performance one can introduce error rate:

$$E_k = \frac{\#(CR_k^S \cap R_k^T) + \#(R_k^S \cap CR_k^T)}{\#D} \in [0..1].$$

It measures the quota of errors; therefore, low values are desirable.

During the evaluation one should collect measurements for each single experiment – query input. The task of the systems comparison based on such observations is not feasible. Thus, the aggregation of the measurements is needed. One approach is based on acquiring individual measurement for each experiment and then averaging over experiments. Or, the measurement might be computed globally over all experiments. The former way is called macro-averaging and the latter way is called micro-averaging. In this paper we use the approach of macro averaging of measurements over experiments with further aggregation of results through statistical tests for acquiring the significance level of the evaluated parameter.

## 8. STATISTICAL ANALYSIS

In this section we will discuss approaches of statistical analysis of collected measurements. Following, there are two approaches proposed. First aims to select the best out of two systems. These can be different systems or the same system being evaluated in different configurations. The second approach is a test, which is designed to check the level of difference significance in the effectiveness of two systems. This test is known as t-Test [20].

Further statistical methods can be applied to all proposed measurements. The interpretation of results, however, depends on the comparison order of specific measurements.

### 8.1 Assumptions

Prior to starting statistical manipulations on retrieved measurements we would like to make some assumptions on the nature of the data processed. These assumptions allow performing further statistical analysis and accepting the results.

The key assumption made is such that measurement values of information retrieval system are normally distributed. On the one hand this assumption is unrealistic; however it follows immediately from the way these measurements are collected. The observed precision, recall, F, and other measures are obtained through averaging measurements obtained by different people. Following, the Central Limit Theorem comes into play.

Central Limit Theorem is one of the most important theorems in probability theory and statistics. It assigns the normal distribution an outstanding role.

Be  $(X_i)_{i \in N}$  a sequence of independent and identically distributed (*iid*) random variables with  $\mu = E[Y_i] \in (-\infty, \infty)$  and variance  $\sigma^2 = Var[Y_1] < \infty$ . Be

$$S_n = \frac{1}{n} \sum_{i=1}^n X_i \quad Z_n = \frac{S_n - \mu}{\sqrt{\sigma^2/n}}$$

and

$$F_n(t) = \Pr[Z_n \leq t]$$

be the distribution function of  $Z_n$ , then

$$F_n \xrightarrow{D} \Phi$$

as  $n \rightarrow \infty$ . Here  $\Phi$  is the distribution function of a normal distribution with mean value 0 and variance 1. Here the convergence in distributions is meant. Therefore, we can assume our measurement as being a realization of a random variable that has normal distribution with unknown parameters  $\mu$  and  $\sigma^2$ .

## 8.2 Comparing Two Systems

Finally we get to the point where collected measurements data will be aggregated and concrete statistical reasoning will be made based on the retrieved results. In this section we will provide approach to comparing information retrieval systems. However, we will restrict our method to the case of comparing two systems. As already mentioned, these can be completely different systems or the same system evaluated under different internal configurations.

Here we use the above stated assumption of observed measurement being *iid* for the both systems. Further, be  $\mu_1$  and  $\mu_2$  respectively the true expectation for systems A and B.

The key idea of this comparison method is to find a confidence interval  $I_\alpha$  for the difference  $\mu := \mu_1 - \mu_2$  and a given confidence level  $\alpha$ .  $\mu$  can be estimated through  $\hat{\mu}$  (mean value). Following in the Figure 4 possible outcomes for  $\hat{\mu}$  are visualized.

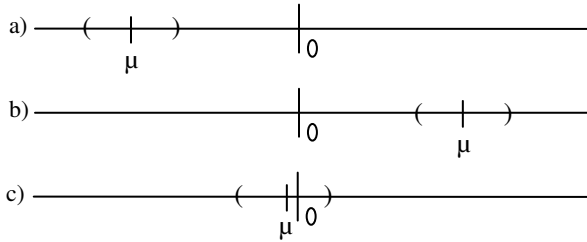


Figure 4. Possible outcomes for  $\mu$ .

There are three different outcomes when comparing two systems for differences in their mean values of the observed parameter. Outcome (a) corresponds to the case when system B is significantly better than system A, while outcome (b) will mean system A be significantly better than system B. In case of outcome (c) clear distinction is not possible. In this case further measurements need to be made to decrease confidence interval to the point it does not contain 0.

To reconstruct confidence interval we will use Paired-*t* confidence interval. To the already discussed assumptions one should add the one that the number of replications (experiments) is exactly the same for both systems. With these assumptions the random variables

$$Z_i = X_i - Y_i$$

with observations  $z_i = x_i - y_i$  are *iid*. This allows us to compute:

$$\hat{z}(n) = \frac{1}{n} \sum_{i=1}^n z_i, \quad S^2(n) = \frac{1}{n-1} \sum_{i=1}^n (z_i - \hat{z}(n))^2$$

as unbiased estimations of  $E[Z]$  and  $Var[Z]$ . Now we can construct  $\alpha$  level confidence interval as:

$$\left[ \hat{z}(n) - t_{n-1, 1-\frac{\alpha}{2}} \sqrt{\frac{S^2(n)}{n}}, \hat{z}(n) + t_{n-1, 1-\frac{\alpha}{2}} \sqrt{\frac{S^2(n)}{n}} \right]$$

In case the number of observations differs for both systems and we would not like to drop the observations in order to analyze equal test data collections we can use modified procedure for confidence interval construction. This procedure is known as Welch Procedure [15]. It assumes that  $n_1$  needs not be equal to  $n_2$ .

Be:

$$\hat{x}(n_1) = \frac{1}{n_1} \sum_{i=1}^{n_1} x_i, \quad \hat{y}(n_2) = \frac{1}{n_2} \sum_{i=1}^{n_2} y_i$$

$$S_x^2(n_1) = \frac{1}{n_1-1} \sum_{i=1}^{n_1} (x_i - \hat{x}(n_1))^2$$

$$S_y^2(n_2) = \frac{1}{n_2-1} \sum_{i=1}^{n_2} (y_i - \hat{y}(n_2))^2$$

Then we compute the estimated degree of freedom as:

$$\hat{f} = \frac{\left( \frac{S_x^2(n_1)}{n_1} + \frac{S_y^2(n_2)}{n_2} \right)^2}{\frac{(S_x^2(n_1))^2}{n_1^2(n_1-1)} + \frac{(S_y^2(n_2))^2}{n_2^2(n_2-1)}}$$

Finally we compute the confidence interval for  $\hat{\mu} = \hat{x}(n_1) - \hat{y}(n_2)$  as:

$$u_{u,l} = \hat{\mu} \pm t_{\hat{f}, 1-\frac{\alpha}{2}} \sqrt{\frac{S_x^2(n_1)}{n_1} + \frac{S_y^2(n_2)}{n_2}}$$

$\hat{f}$  is usually not an integer, therefore one should look at neighbored integers when finding  $t_{\hat{f}, 1-\frac{\alpha}{2}}$  and pick the larger one.

### 8.3 Significance Test (t-Test)

In the previous section we have discussed approach for comparing two systems. This approach can be extended to the case of determining the level to which system A outperforms system B. This kind of analysis can be performed with the help of statistical test, and to be more precise, the t-Test.

The evaluation that was discussed in the previous section should be continued to the point when the confidence intervals for the observed parameter for both systems are not overlapping. This means that there is no need to perform all the experiments (feed in all the queries), but continue evaluation to the point when clear distinction can be made. However, the number of queries we possess might not be sufficient for the experiment to deliver any results. This might be the case when using proposed in section 4.2 test collections. In this case one might use the approach proposed in this section. It will always deliver results, but again the level of significance will depend on the amount of information processed in the experiment.

Again, we have observations of the tested measurement for both systems.  $X_i$  for the system A and  $Y_j$  for the system B. As we have explained in the assumptions section:

$$X_i \sim N(\mu_x, \sigma_x^2) \quad Y_j \sim N(\mu_y, \sigma_y^2)$$

This gives us a possibility to construct a test. The null hypothesis is  $H : \mu_x - \mu_y \leq d_0$ , or the difference of expectations for the observed parameter is less than some tested value  $d_0$ . The alternative hypothesis is that this difference is larger than tested value  $d_0$  meaning that system A performs for this parameter better than system B -  $K : \mu_x - \mu_y > d_0$ . Therefore, the aim of the test is to reject  $H$  and to confirm  $K$  to some significance level.

In t-Test we reject  $H$  when:

$$T > t_{n_1+n_2-2; 1-\alpha} \quad (*)$$

where  $n_1$  is the number of experiments for system A and  $n_2$  is respectively the number of experiments for system B. In most cases  $n_1 = n_2 = n$ . However, they need not be equal.  $\alpha$  is the significance level to which we test  $d_0$ .  $t$  is the quantile of the student-t distribution and  $T$  is our test statistics which is obtained as:

$$T = \frac{\bar{X} - \bar{Y} - d_0}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \hat{\sigma}^2}}$$

Where  $\bar{X}$  and  $\bar{Y}$  are the arithmetic averages of the observed parameter for both systems. And  $\hat{\sigma}$  is the standard quadratic failure for the random variable  $X_i - Y_i$ . Because  $T$  is

calculated using random variables,  $T$  itself is a random variable. It can be shown that  $T$  has a student-t distribution:

$$T \sim t_{n_1+n_2-2}$$

where  $n_1 + n_2 - 2$  is the degree of freedom used in the student-t distribution. The last thing we need in order to be able to perform all the calculations is the method of calculating  $\hat{\sigma}^2$ .  $\hat{\sigma}^2$  is the estimator of the variance for the random variable  $X_i - Y_i$  and can be obtained as:

$$\hat{\sigma}^2 = \frac{(n_1 - 1)S_x^2 + (n_2 - 1)S_y^2}{n_1 + n_2 - 2}$$

where  $S_x^2$  and  $S_y^2$  are the estimators of the variance of the random variables  $X_i$  and  $Y_j$ . The formula to obtain estimator of the variance is provided in the previous section.

First of all with this approach experimenter is able to vary the parameter  $d_0$  – the level to which system A outperforms system B in the evaluated parameter and the  $\alpha$  level. Also, it is possible to get the corresponding  $\alpha$  level for the tested data by calculating test statistics and solving inequality (\*).

## 9. CONCLUSIONS

There is no such thing as a watertight method for evaluating an information retrieval system [1]. Any existing approach to evaluating or comparing information retrieval systems will have to deal with heuristics to some extent only for the reason of this process been highly dependant on human factor. In this paper we discussed the approach to conducting evaluation of information retrieval system starting from preparation of workload, conducting experiment and finally statistical data analysis. This approach is suitable for comparison of two information retrieval models or evaluation of a single system under different configurations of the model used. In this paper we proposed the approach to study of eTVSM. Through reconfiguration of eTVSM and following re-evaluation of the model it is aimed to come up with the optimal model configuration.

The work reported in this paper can be treated as the set of instructions to take in order to perform quantitative evaluation of any information retrieval system. By strictly following this instructions one will be able to evaluate eTVSM or any other model. It is also possible to reuse some of the parts of the proposed approach or extend it to suit specific requirements. However, theoretical value of this work is in its completeness, thus future pure realization of the proposed approach is highly encouraged.

Researchers in the information retrieval field have devoted a significant amount of time in developing good, standardized evaluation techniques. This work should be treated as the standardized approach to eTVSM evaluation which subsumes the

technique suitable for the typical information retrieval model evaluation.

## 10. FUTURE WORK

This paper is theoretical base for practical evaluation. The future work, therefore, will have to deal with practical realization of the proposed approach to information retrieval model evaluation and study of eTVSM. This includes preparation of test collection and reuse of the existing ones, conducting experiment, taking measurements and further statistical reasoning on collected measurements. For the eTVSM not only the statistical evaluation with further selection of the best configuration should be made, but a complete approach to optimal configuration of eTVSM should be developed.

## 11. ACKNOWLEDGMENTS

I would like to thank to Dr. rer. pol. Dominik Kuroпка for advices in the field of information retrieval, PD Prof. Dr. Hannelore Liero for consultations in the statistics field, and all the participants of Information Filtering and Retrieval seminar for giving valuable comments on this work: Sergey Smirnov, Lars Trieloff, Raphael Audet and Rémi Liance.

## 12. REFERENCES

- [1] Jones, K.S. (Hrsg.): Information Retrieval Experiment. Butterworth, 1981.
- [2] Yang, Y.; Liu, X.: A re-examination of the text categorization methods. In Hearst, F.G.; Tong, R. (Hrsg.): Proceeding of the 22<sup>nd</sup> Annual ACM SIGIR Conference on Research and Development in Information Retrieval. Berkley, 1999.
- [3] Text Retrieval Conference (TREC): <http://trec.nist.gov>.
- [4] Sanderson M.; Zobel J.: Information Retrieval System Evaluation: Effort, Sensitivity, and Reliability. SIGIR'05, August 15-19, 2005.
- [5] Information Retrieval Special Interest Group (ACM SIGIR): <http://www.acm.org/sigir/>.
- [6] Sanderson M.; Joho. H.: Forming Test Collections with No System Pooling. SIGIR'04, July 25-29, 2004.
- [7] Soboroff I.; Robertson S.: Building a Filtering Test Collection for TREC 2002. SIGIR'03, July 28-August 1, 2003.
- [8] Cormark G.V.; Palmer C.R.; Clarke C.L.A.: Efficient Construction of Large Test Collections. SIGIR'98, 1998.
- [9] Sanderson M.: Reuters test collection. Glasgow University Computing Science Department. 11 June 1994.
- [10] Reuters Corpora: <http://groups.yahoo.com/group/ReutersCorpora/>
- [11] National Institute of Informatics (NTCIR Project): <http://research.nii.ac.jp/ntcir/permission/perm-en.html>
- [12] Blair D.C.; Maron M.E.: "An evaluation of retrieval effectiveness for a full text document retrieval system". Communications of the ACM, Vol. 28, Num. 3, Pages 289-299, 1985.
- [13] C.J. van Rijsbergen. Information Retrieval. Butterworth, London, 1979.
- [14] Heinz Bauer. Wahrscheinlichkeitstheorie und Grundzüge der Masstheorie. Walter de Gruyter, Berlin, 1968.
- [15] Christos G. Cassandras. Discrete Event Systems - Modeling and Performance Analysis. Aksen Associates, Boston, 1993.
- [16] Christos G. Cassandras and Stephane Lafortune. Introduction to Discrete Event Systems. Kluwer Academic Publishers, Boston, 1999.
- [17] William Feller. An Introduction to Probability Theory and Its Applications - Volume II. John Wiley, New York, second edition, 1968.
- [18] John L. Hennessy and David A. Patterson. Computer Architecture - A Quantitative Approach. Morgan Kaufmann, Amsterdam, Boston, third edition, 2003.
- [19] Raj Jain. The Art of Computer Systems Performance Analysis - Techniques for Experimental Design, Measurement, Simulation, and Modeling. Wiley Professional Computing. John Wiley and Sons, New York, Chichester, 1991.
- [20] Roxy Peck, Chris Olsen, and Jay L. DeVore, editors. Introduction to Statistics and Data Analysis. Duxbury, 2000.
- [21] Kuroпка, D.: Modelle zur Repräsentation natürlichsprachlicher Dokumente - Information-Filtering und -Retrieval mit relationalen Datenbanken. Logos Verlag, Berlin, 2004.
- [22] Robertson, S.: The methodology of information retrieval experiment. In Jones, S. (Hrsg.): Information Retrieval Experiment. Butterworths, 1981.
- [23] K. J. Hastings: Probability and Statistics Addison-Wesley 1997.
- [24] H. Lauter, R. Pincus: Mathematisch- Statistische Datenanalyse Akademie-Verlag 1989.
- [25] Reuters-21578: <http://www.daviddlewis.com/resources/testcollections/reuters21578/>.
- [26] Glasgow IDOM test collections: [http://www.dcs.gla.ac.uk/idom/ir\\_resources/test\\_collections/](http://www.dcs.gla.ac.uk/idom/ir_resources/test_collections/)