Relevance & Assessment: Cognitively Motivated Approach toward Assessor-Centric Query-Topic Relevance Model

Bassam Haddad

University of Petra, Department of Computer Science, Amman-Jordan
haddad@uop.edu.jo

Abstract: This paper intends to introduce a novel model for query-topic relevance assessment from assessor and cognitive point of view in the sense that relevance is a multidimensional cognitive and dynamic conception. The focus of this presentation is concentrated on modeling the concept "Query Associative Vocabulary of Relevance" to emphasize the value of integrating intuitive, descriptive, multi-valued assessment, and agreement in the process of creating a query-topic relevance data. As this model differentiates between different types of query-topics and levels of relevance, it provides a facility to enhance the quality of relevance data by re-evaluating the resulted associative vocabulary at each cycle of refinement. This aspect is of importance, as it is directed toward extracting as much advantage from human assessment as possible. A prototype of this model has generated in an initial run a relevance dataset of 20.710 relevance assessor's feedback and a co-occurrence matrix of 39607 terms distributed in intuitive, descriptive and document associative vocabularies. Most of the assessor feedback is descriptive produced by humans in context of establishing a relevance relationship between a query-topic and related documents. Furthermore, classifying query relevance datasets according to grades of agreements among judgments is useful as it gives a better overview of the performance of the considered system and the comparison of different datasets in context of consistency and performance becoming easier. Despite the importance of relevance in designing and evaluating Information Retrieval Systems as possible inter-cognitive systems, a consensus on definition is still debatable. However, considering relevance as a multidimensional cognitive and dynamic conception provides researcher with a research track to evaluate the performance of interactive and inter-cognitive processes in terms of the multidimensionality and cognitive aspects of relevance.

Keywords: Relevance Assessment; Query-Topic Modelling; Relevance Dataset; Assessor-Centric; Judgments Agreements; Cognitive Linguistics; Information Retrieval, Search Engine Performance; Word Associative Network, Cognitive InfoCommunication; Topic Model

1 Introduction and Motivations

Relevance is still a critical issue of Information and Cognitive Science. Despite its significance in designing and evaluating Information Retrieval Systems [12]; in particular in context of employing them within inter-cognitive processes, a consensus on definition is still debatable. In the literature, relevance can be considered from different perspectives: from *the system (topicality matching), user satisfaction and relevance-feedback, multidimensionality of topicality utility* and from the *cognitive perceptive* [21].

However, this presentation proceeds from an *assessor-oriented model* considering *the cognitive aspect and the multidimensionality of relevance* in the sense; it is considered as a *multidimensional cognitive and dynamic conception*.

On the hand, a central question is still controversial: *How does an assessor conceive a document as relevant?* The vagueness involving its nature led to confusion in finding proper criteria for representation and assessment. The process of relevance assessment enforces human brain to highest concentration and activity, whereas *intuitive background* of the assessors within an inter-cognitive communication [3] might affect the quality of a processing of relevant information. According to [18], relevance judgment is inconsistent; it can be affected by 40 and even according [22] by 80 factors. For Example, the following factors might affect the relevance assessment:

- **On the Assessor Level***: cognitive style, bias, education, intelligence and experience, motivation, etc.*
- **On Information Request and Need Level**; *i.e. query-topic formulation: difficulty, subject and textual features, query type (one term, structured, unstructured), multimedia features, etc.*
- **On Document Level:** *precision, difficulty, importance, novelty, aboutness, aesthetics.*
- **Assessment Conditions:** *size of the document set, Time for judgments, experiential environment, interaction modality, visualization, etc.*
- **Assessments Type and Information System:** *binary, multi-valued, descriptive assessment, system access, relevance modelling, etc.*

Correspondingly, [13] formalized similarly this aspect by emphasizing, that there are many kinds of relevance, and not just one, which can be represented by *four formal dimensional space*; i.e. *Information resource*; e.g. documents, *requested need*; e.g. query or topic representation and *assessor's condition*, and *background knowledge* are the major factors involved in the relevance assessment process.

Different relevance sets of relevance assessment might be observed under different judgment's conditions; such as *assessor's motivation, assessor*

experience or the *intuitive knowledge* of the used topics. Furthermore, despite the closeness between the relationship between relevance assessment and relevance feed-back concept, this work distinguishes between these terms, in sense that goal of relevance assessment is to provide a *reference of relevance* for measuring performance of an Information Retrieval, which might be integrated within an CogInfo-Communication process, while relevance-feedback is focused toward improving the precision by evaluating and reformation and expansion user's feed-back (User-Satisfaction model of Relevance).

The process of creating a traditional relevance corpus in TREC for instance, seems to be not visible from a cognitive point of view specially in the case of considering multiple assessments for different documents. The overall intuitive vocabulary of the assessors and even the inspected document vocabulary are not visible in the process of assessment. TREC relevance assessment relies strongly on the pooling principle and a batch processing evaluation. The assessors are responsible for formulating and at the same time for the relevance assessment, whereas their *overall multidimensional intuitive background* of the investigated topics is not considered in the assessment process. Topic and document terms possibly with cognitive phonetic spell errors, polysemous terms or informal content [7], [9] confuse the inter-cognitive process of assessment. Some assessors might consider, due to a possible cognitive load, irrelevant or marginally relevant documents as relevant and even highly relevant. Such kind of miss-communication in the process of relevance assessment can be considered as a kind of misinterpretation and a disturbing factor for creating a representative relevance data. Topic terms and their *intuitive associative network*, *documents vocabulary* and even *human-machine interaction* might affect this process. In this context, considering the overall intuitive or the *cognitive vocabulary* generated by different assessors provide us with a valuable re-usable source for topic *reformulation* and *assessment*. This paper will stress therefore on capturing this aspect when creating a relevance data. This implies the attempt to formalize the overall intuitive vocabulary of multiple assessors involved in a relevance assessment experiment, representing multidimensional assessor's views of an assessment.

Furthermore, TREC evaluation methodology is predominantly based on the binary logic of relevance, i.e. dichotomous judgment such as relevant or not-relevant judgment. Despite the overall relative stability of TREC based retrieval performance [20], there are still some critics coming from the lack of practicality; i.e. *the utility dimension*, and the potential meaning and usefulness of a retrieved document to the user in context of measuring the performance. This issue might be supported in connection with the increasing demand of finding *highly relevant documents* expressed in terms of degrees of document relevance. For example, binary assessments allow the assessor to classify *marginally relevant* and *highly and even very highly relevant* document to the same relevance class. However, in the meantime there are several TREC web tracks utilizing points-based relevance scale (not-relevant, relevant, highly relevant) [10], [11].

In the view of this presentation, relevance assessment should be assessor-based, requiring some *dynamic cycle of refinement* and *ratification* under considering appropriate preprocessing steps to simplify a possible inter-cognitive communication. In this context, this approach is differentiating between variant types or levels of relevance depending on the depth of refinement. The depth of refinement relies dominantly on three major aspects: *relevance assessment*, *assessor feedback* and *agreement*; whereas the grades of agreement should be considered at each level of assessment. And finally, the overall "*Vocabulary of Relevance"* created during the relevance assessment should also be captured and formalized as reference for any further refinement. The last aspect represents a core constituent of the proposed model; as the resulted "*Vocabulary of Relevance"* might make Data of relevance more visible and reusable for IR-Systems evaluation.

In the proposed approach, datasets of relevance are represented as "Associative *Vocabularies" depending* on depth of the captured assessor's initial vocabulary *before* and *after* each assessment feedback. At each level of assessment, the *priming principal* can be utilized to capture the intuitive assessor's vocabulary for each query-topic, whereas after an assessment the assessors are invited to create new query for each already assessed document. The resulted queries are then subject to an overall multi-valued assessor agreement to estimate the consistency between a group of judges, and to use as measure for relevance.

In the approach, the process of relevance assessment can be regarded as *cognitive process* of establishing a *relevance relationship* between query-topic *latent words and documents associative networks;* see Figure 1. Adopting this approach requires developing an *assessor-oriented interactive assessment system* considering some kind of an *inter-cognitive communication,* assessor's *relevance feedback* and *judgment-agreement.* For implementing such a system, the priming principle has been utilized for creating initial intuitive term or *word-associative network* of investigated queries-topics. These *associative networks* can act as an initial human-based "*Query Associative Vocabulary"*. For generating a useful *human-machine document-topic* related vocabulary, the priming principle can also be utilized for establishing *document-topic relationship* by requesting assessors reading some documents and describing their topics in their words. This *intuitive-machine influenced Vocabulary*, contains implicitly a useful relevance assessment, which might be used in query formulation and further assessment [7]. These *associative document-topic* relationships can act as an initial human-document-associative vocabulary.

Finally, assessors are requested to assess the relevance in the traditional way, however under consideration a *non-binary*; i.e. *non-dichotomous judgment* and an agreement of the multiple judgments. Furthermore, software engineering aspects such as reusability, flexibility and others should also be considered in creating targeted Relevance Vocabulary [6].

For testing the resulted system an Arabic Corpus[1] has been considered containing 110 Query-Topics and 3300 documents extracted from the ClueWeb [8].

1.1 Related Work

As mentioned earlier, most work concerning creating relevance corpora relies dominantly on TREC tracks. The traditional process of creating relevance corpora in TREC has not been significantly changed. It is based on the pooling principle to ensure the retrieved collection of documents is comprehensive as possible and batch processing evaluation. However, in context of using many-valued logic for relevance assessment, there are in the meantime, some papers reporting on the increasing demand for considering multiple-point assessment. [16] reassessed TREC documents pools on 38 Topics to build a sub-corpus of highly relevant documents based on the four-point scale. He found 39% agreement with the TREC relevance assessment.

In connection to the meaning of the Human-Machine Interaction and user-based evaluation in establishing a relevance assessment, there is also related work. Turpin and Scholar [19] stressed on the weak co-relationship between user performances against precision-based measures of Informational Retrieval. In this context, a precision-based user task measured by the time needed to identify a relevant document and a recall-based task measured by the number of finding relevant documents within a determined period of time. They observed 45% agreement with TREC relevance. [2] Found even 65% agreements with the official TREC judgments in an Interactive IR experiment.

Furthermore, in context of measuring the consistency of the agreement among relevance judges, there is some similarity between this approach and research presented in [14] and [22]. However, missing judge's assessments were considered. Moreover, this approach has tried to deviate from the traditional *kappa agreement notion*, as our approach is heavily considering non-binary judges assessment, besides the critics on this approach [17].

In context of Arabic script-based corpus evaluation [1], most studies rely strongly on the TREC 2001/2002 cross-language retrievals track [4]. In this track, based on collaborative work of different teams, 5909 documents over 50 topics were found to be relevant with 118 relevant documents per topic after considering total of 41 runs on an Arabic Corpus of 383,872 documents [5]. The topics were originally prepared in *English and then translated into Arabic*. Unlike the proposed approach, the traditional TREC Approach for relevance assessment was binary

[1]In spite of fact that the proposed model for relevance assessment is *language independent*, the selection of Arabic came from a pragmatics point of view related to researcher current affiliation and research in context of creating Cognitively-Motivated Query Abstraction Model [7], [8] and [9].

(Yes/No). However, there is some recent research concerned with optimizing retrieval of informal content of Arabic (such as Dialect or non-lexical terms) [15].

The remaining parts of the paper will be focused on modeling Vocabularies of Relevance; particularly on introducing the concept "Query Associative Vocabulary of Relevance" and "Assessors Agreement on Relevance".

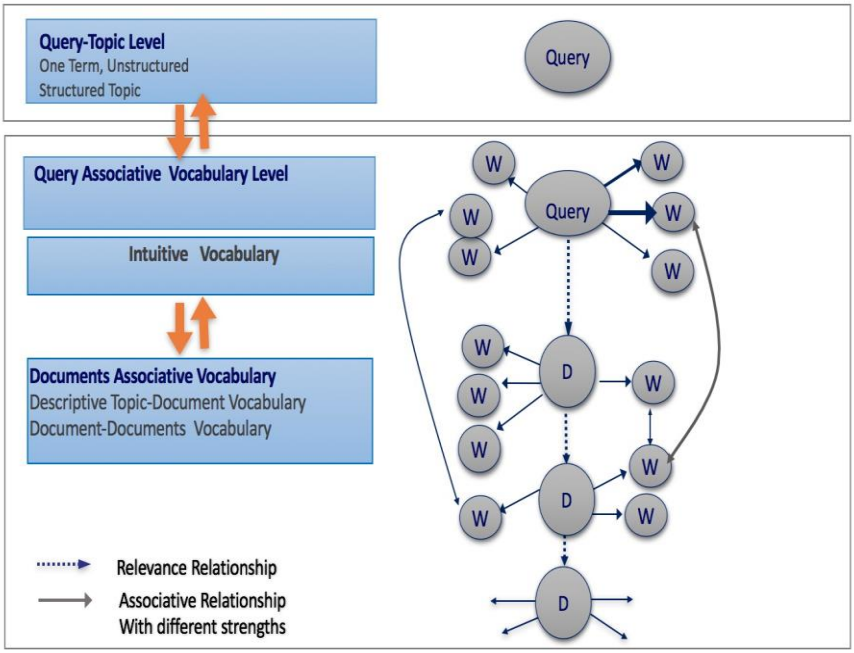

Figure 1
Query-Topic Associative Levels considering Intuitive, Descriptive and Document Associative Vocabulary

2 Modeling Vocabulary of Relevance

A traditional test collection consists usually of:

• *Set of Topics*

• *A Set of Related Documents.*

• *Relevance judgments correlating query-topics to certain documents.*

However, the proposed approach will elaborate on the interrelationship between these sets from a *cognitive point of view* focusing on the role of the assessors for establishing relevance relationship between queries related documents. Therefore,

this approach can be regarded as assessor-based and cognitively oriented. Furthermore, it aims at making a relevance assessment visible and consistent among the assessors by capturing instances of assessor's vocabularies at different levels of depth and refinement. As an assessor has to assess the relevance of a query-topic in context of a text-document based on its words, his *background-vocabulary plays* a decisive role in establishing a relevance relationship between a topic and a text-document. In this presentation, the dimension "*intuitive"* and/or "*associative"* vocabulary will be used in context of *Productive*[2] *and Receptive Vocabularies*[3]. Furthermore, this presentation will differentiate between two major concepts:

- *Query Associative Vocabulary of Relevance* ***(QAV)***
- *Query Datasets of Relevance* ***(Q-Rel-Set)***

A Query Associative Vocabulary of Relevance can be viewed as associative word-networks reflecting assessors *intuitive* and document *associative* background knowledge, while Query Relevance Datasets represent the results of the process of establishing a relevance relationship between queries and related documents. In this context, a query-topic is not considered only through its terms, but rather more through an Associative Word-Network[4] capturing a query-topic intuitive and document associative network. Furthermore, the process of Relevance assessment is considered as an abstract process of establishing a relevance relationship between a Query Associative Relevance Vocabulary; i.e. query associative word-networks and documents associative networks, see Figure 1.

To formalize these aspects, some preliminary definitions will be introduced.

2.1 Preliminary Notation

Let

• D=$\{ d_1, d_2, ..., d_n \}$ be the set of all considered documents.

• J =$\{ J_1, J_2, ..., J_m \}$ be the set of judges, who should perform the relevance assessment.

• Q=$\{ q_1, q_2, ..., q_q \}$ be set of considered queries-topics.

[2]Productive Vocabulary is declared to be the set of words that can be produced by assessors within an appropriate context of relevance.
[3]Respective Vocabulary is specified to be the set of words understood by assessors when heard or read or seen forming a human vocabulary.
[4] A Query-Topic based Associative Network represents a latent structure of the related Topic.

Furthermore, the queries are classified in the following structural types:

- **Query Type-I: One Term Query-Topic**. Defined as the class of topics consisting of one term query. A query of type one is denoted by $q_{i\langle I\rangle}$; e.g.:

  $q_{i_1\langle I\rangle} = \langle Education \rangle$, $q_{i_2\langle I\rangle} = \langle Energy \rangle$ and $q_{i_3\langle I\rangle} = \langle Cells \rangle$.

- **Query Type-II: Unstructured Query-Topic**. Defined as the class of queries represented in unstructured from. A query of this type is represented by multiple related words or terms, however not structured from. E.g.:

  $q_{i_1\langle II\rangle} = \langle Game, Internet, Programs \rangle$

  $q_{i_2\langle II\rangle} = \langle Surgery, Heart, Operations \rangle$.

- **Query Type-III: Structured Query-Topic**. Defined as the class of queries representing a query in a structured form. This type represents query in the usual form; e.g.:

  $q_{i_1\langle III\rangle} = \langle Real\ \ Estates\ in\ United\ Arab\ Emirates \rangle$

  $q_{i\,3\langle III\rangle} = \langle When\ can\ the\ lender\ hold\ the\ proerties\ back? \rangle$

Furthermore, the following applicative functions are denoted as follows:

- $\langle q_i\ \langle \mathrm{D} \rangle \rangle$ denotes a vector of documents, which are associated with the query q_i and can be extracted based on some search strategy; e.g.:

  $\langle q_i\ \langle \mathrm{D} \rangle \rangle = \langle d_1, d_2, ..., d_l \rangle$ and $\langle q_i\ \langle d_j \rangle \rangle = \langle d_j \rangle$ represents the j-document in $\langle q_i\ \langle \mathrm{D} \rangle \rangle$.

- $\langle q_{i\,INT} \rangle_{\mathrm{J}}$ be an instance of the Intuitive Vocabulary of the query q_i, which is associated with a group of assessors J and can be created by capturing the priming effect of the query q_i. Analog $\langle q_i \rangle_j$ represents priming effect of the query q_i, by some judge $j \in \mathrm{J}$.

- $\langle q_i \langle \mathrm{D} \rangle_{DIS} \rangle_{\mathrm{J}}$ be an instance of the Descriptive Relevance Assessment Vocabulary produced by the group J for the query q_i when observing the documents $\langle q_i\ \langle \mathrm{D} \rangle \rangle = \langle d_1, d_2, ..., d_l \rangle$.

- $\left\langle q_i \left\langle \boldsymbol{d} \right\rangle_W \right\rangle_{\mathrm{J}} = \left\langle w_1, w_2, ..., w_m \right\rangle_{\mathrm{J}}$ with $w_i \in [0,1]$ be a vector from the space of weighted assessments associated with the document $d \in \mathrm{D}$ in context of establishing a relevance relationship with the observed query q_i produced by a set of judges J .

- $\left\langle q_i \left\langle \mathrm{D} \right\rangle_W \right\rangle_{J_k} = \left\langle w_1, w_2, ..., w_l \right\rangle_{J_k}$ with $w_i \in [0,1]$ be a vector of weighted assessments associated with documents $\left\langle q_i \left\langle \mathrm{D} \right\rangle \right\rangle = \left\langle d_1, d_2, ..., d_l \right\rangle$ in context of establishing a relevance relationship with the observed query q_i produced by some judge $J_k \in \mathrm{J}$.

2.2 Query Associative Vocabulary of Relevance

A Query Associative Vocabulary can be viewed as an Associative Query-Network, which might be used in an assessment process. Capturing such associative Vocabulary is difficult to determine. However, this approach proposes proceeding from an initial instance for such Vocabulary, which might be augmented and refined by multiple feedbacks within an agreement strategy. In this presentation, an initial Query-Network is considered in view of the assessors from the flowing points of view:

a. Intuitive assessor's feedback as Query Intuitive Vocabulary **(QIV)**.

b. Productive assessor's feedback as Query-Document Associative Vocabulary.

c. Document Associative Vocabulary, (**DAV);** see Figure 2.

The associative vocabularies in (a) and (b) represent possible instances of Assessors Productive Query Vocabulary in context of *intuitive* and *descriptive* abilities of the assessor, while Associative Document Vocabulary in (c), represents a document associative network, which might be estimated by classical n-gram analysis. However, the focus of this presentation will be on modeling of Assessors Associative Vocabulary of Query. In this presentation the assessor's feedback in (a) and (b) will be considered as *Query Associative Vocabulary* **(QAV).**

It is clear that an assessor; when establishing a relevance relationship between a query and a document can't consider all aspects of associative relationships. He/ She might express this kind of uncertainty by estimating the relevance relationship relying on many-valued or descriptive and declarative relevance assessments. On the other hand, as capturing the whole types of associative networks; i.e. associative vocabularies of a topic and document is also not possible, this approach attempts to formalize these under the relativity of these aspects for all

assessors. This view can be implemented through multiple inter-cognitive communications, before and after having more relevance details at different sessions of communication. This view implies for example, to estimate the Assessor Intuitive Vocabulary by capturing the priming-effects of all involved assessors. Furthermore, topic associative vocabulary can be estimated based on the agreement among all assessors and their feedback in the form of creating of reformulating the initial query relying on more details after exploring the related document and even its meta-data. Each captured associative word-network should be subject of selection and agreement of involved assessors. QAV is proposed to be estimated over assessor's productive vocabulary, on the following levels of observations and refinements:

- **Productive Effect Level;** i.e. when reading or seeing or hearing a query-topic independent of a document. This *dimension of relevance* is concerned with representing the basic contextual relevance of query as an instance of the associative network for a query. Instances of a query associative network can be generated by considering query associated word delivered by assessors before starting an assessment process. In other words, it aims at capturing the priming-effect of a topic for all assessors. For Example, relying on certain J assessors, the query $\langle Cells \rangle$ has produced on the initial run of the experiment the following intuitive Effect:

$$\langle Cells_{INT} \rangle_{J} = \left\langle \begin{matrix} Beehives, Stem, Blood, Biology, Body, Human, \\ Solar, Terrorist, Nerve, ... \end{matrix} \right\rangle \quad (1)$$

  with different frequencies. Such query associative set can be viewed as weighted associative word-network reflecting the most associative words with query-topic.

- **Active Productive Level;** i.e. Relevance based on judges-agreement, when describing a relevance relationship between a topic relying on assessor's receptive vocabulary. E.g. after observing or reading a query description, document words, *and/or Meta terms of some document*. In this context, this approach differentiates between two basic kinds of associative vocabularies of Relevance estimating the productive vocabulary in terms of relevance assessments:

  a. Query-based Descriptive Relevance Assessment. This type reflects judges' assessment in term of establishing a relevance relationship between a query and a document by creating or reformulating a query text or topic for a certain document describing a high relevance relationship after reading and having more details of the document. In other words, assessors are requested to answer the question, *what is the best formulation you propose to inquiry the*

investigated document? The influence of document vocabulary and its associative network should play an important role in the assessment process, as the assessor might rely on certain terms occurring in the document. This type of assessment can be considered as query reformulation or expansion, relying on *assessor's receptive vocabulary* of document and on the initial query. For Example, based on J , the document d=ar004-15-28[5] with the query $\langle Cells \rangle$ has produced the following Descriptive Relevance:

$$\left\langle Cells \left\langle d \right\rangle_{DIS} \right\rangle_{J} = \left\langle \begin{matrix} Aids\ Aids\ virus, treatment\ of\ immune\ deficiency, \\ immune\ cells, destruction\ of\ cells, ... \end{matrix} \right\rangle \quad (2)$$

b. Weighted Non-Binary Query Relevance Assessments; this type reflects judge's assessment in term of establishing a numerical relevance relationship between a query and a document after reading document text with more details in the interval $w \in [0,1]$. For example, the responded assessors have evaluated the relevance relationship of the query $\langle Cells \rangle$ to the document d=ar004-15-28 with the following vector:

$$\left\langle Cells \left\langle d \right\rangle_{w} \right\rangle_{J} = \langle 0.75, 0.5, 0, 0.25, 0.75, 0, 1, 0, 0.25, 0.25, 0.25, 1, 0.75, 0.75, 0.25, 0.25 \rangle \quad (3)$$

In the following these ideas will be formalized.

Definition 1 (*Query Associative Vocabulary of Relevance, QAV)*

Let

- $q_i \in \mathrm{Q}$ be a query-topic of some type.
- $\mathrm{J} = \{ J_1, J_2, ..., J_m \}$ be a group of assessors.
- $\langle q_{i\,INT} \rangle_{\mathrm{J}}$ be an instance of the Intuitive Vocabulary of the query q_i, which is associated with a group of assessors J and, can be created by capturing the priming effect of the query q_i. Analogy $\langle q_{i\,INT} \rangle_{\mathrm{J}}$ represents associative effect of the query q_i by some judge $J \in \mathrm{J}$.
- $\left\langle q_i \left\langle \mathrm{D} \right\rangle_{DIS} \right\rangle_{\mathrm{J}}$ be an instance of the Descriptive Relevance Assessment Vocabulary produced by the group J for the query q_i when observing the documents $\langle q_i \langle \mathrm{D} \rangle \rangle$ then:

[5] d=ar004-15-28 is a real document extracted from the ClueWeb2009

(a) An instance of the Associative Vocabulary of the Query $q_i \in \mathrm{Q}$ is estimated by:

$$\left\langle QAV_{q_i} \right\rangle_{\mathrm{J}} \subseteq \left\langle q_{i_{INT}} \right\rangle_{\mathrm{J}} \cup \left\langle q_i \left\langle \mathrm{D} \right\rangle_{DIS} \right\rangle_{\mathrm{J}} \tag{4}$$

(b) An instance of the Associative Vocabulary of all Query-Topics is estimated by

$$\left\langle QAV_{\mathrm{Q}} \right\rangle_{\mathrm{J}} \subseteq \left\langle \mathrm{Q}_{INT} \right\rangle_{\mathrm{J}} \cup \left\langle \mathrm{Q} \left\langle \mathrm{D} \right\rangle_{DIS} \right\rangle_{\mathrm{J}} \tag{5}$$

$\left\langle QAV_{\mathrm{Q}} \right\rangle_{\mathrm{J}}$ represents the space of a global associative word-network of all involved query-topics and their associative word produced by a group of assessors.

Definition 2 *(Query-Topic Relevance Datasets)*

- Let $q \in \mathrm{Q}$ be a query of some type.
- Let $\left\langle q \left\langle D \right\rangle_W \right\rangle_{J_k} = \left\langle w_1, w_2, ..., w_l \right\rangle_{J_k}$ with $w_i \in [0,1]$ be a vector of weighted assessments associated with the documents $\left\langle q \left\langle \mathrm{D} \right\rangle \right\rangle = \left\langle d_1, ..., d_l \right\rangle$ in context of establishing a relevance relationship with the observed query-topic q and produced by some judge $J_k \in \mathrm{J}$. Accordingly

$$\left\langle q \left\langle \mathrm{D} \right\rangle_W \right\rangle_J = \left\langle \left\langle q \left\langle \mathrm{D} \right\rangle_W \right\rangle_{J_1}, \left\langle q \left\langle \mathrm{D} \right\rangle_W \right\rangle_{J_2}, ..., \left\langle q \left\langle \mathrm{D} \right\rangle_W \right\rangle_{J_m} \right\rangle \tag{6}$$

represents all assessments of the all assessors for the query q associated documents such that $\left\langle q \left\langle \mathrm{D} \right\rangle \right\rangle = \left\langle d_1, d_2, ..., d_l \right\rangle$.

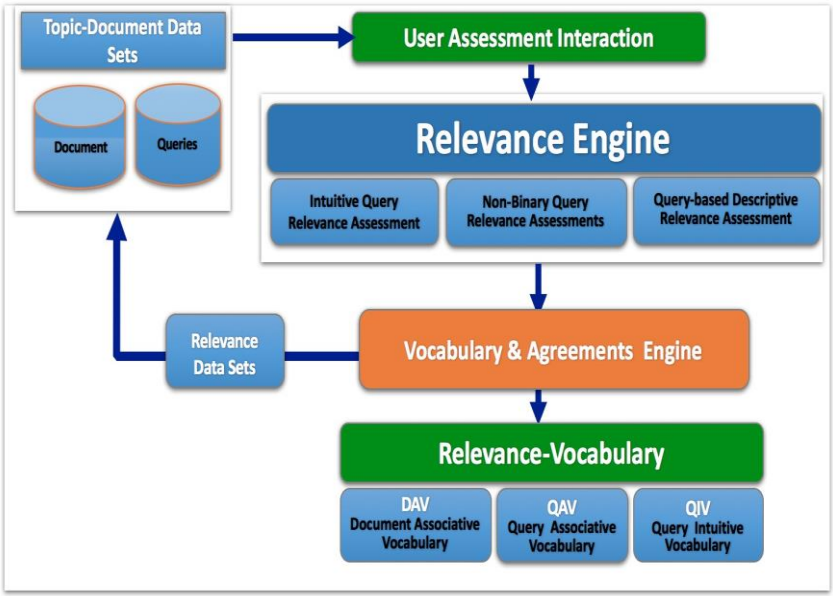

Figure 2
Components of the Proposed Mode

A Relevance Dataset $\langle q\langle \mathrm{D}\rangle\rangle_R$ for the Query q is then defined as the space of a query-topic associated documents and their assessments vectors created by all judges

$$\langle q\langle \mathrm{D}\rangle\rangle_R = \left\langle \langle q\langle \mathrm{D}\rangle\rangle, \langle q\langle \mathrm{D}\rangle_W\rangle_J \right\rangle \tag{7}$$

2.3 Model Architecture

As mentioned above, relevance assessment should be focused on the assessor. And, it requires some cycle of refinement and ratification under considering suitable preprocessing steps to simplify the assessment communications. This approach differentiates furthermore between variant types or dimensions of relevance depending on the depth of refinement. The depth of refinement relies dominantly on three major aspects relevance assessment, assessor feedback and agreement. In addition, intuitive, descriptive and many-valued or multiple relevance assessments were proposed at each level of assessment. The overall vocabulary of Relevance created during the relevance assessment should also be captured and formalized as reference for any further refinement. This last aspect represents a core constituent of the proposed model; as the resulted vocabulary of Relevance might make data sets of relevance more visible for IR-Systems relying on it by evaluation.

Based on the above motivations and definitions, this presentation proposes the following Architecture, which has been implemented[6] and utilized in creating an initial dataset of relevance. The Architecture has two major components, see Figure 2:

- **Relevance Engine.** Based on user interactive assessments capturing the overall intuitive word-network of different query types, query descriptive and many-valued Relevance assessments, the Relevance Engine prepares data networks to creating Relevance Vocabularies.
- **Vocabulary Engine.** Data-Networks will be converted to initial Relevance Datasets to be subject to further assessors-based refinement satisfying some stable grade of overall agreement of consistency. At this step, Query Intuitive, Associative and Document Associative Vocabularies will be created.

3 Grades of Agreement and Disagreement

Relevance datasets consist of collections of relevance relationships organized according to some specific topics or queries to certain related documents. The grade of relevance of some query for some certain documents is captured through assessment registered by multiple judges. As mentioned earlier, human judgment might be subject to different factors, which might affect the outcome of relevance datasets such as judge background, document type, judgment conditions and type of the query.

The focus of attention of this presentation was till now on modeling a *"Query Associative Vocabulary of Relevance"*, to stress on the value of intuitive and descriptive relevance and non-binary assessment. However, the essence of creating a stable dataset of Relevance needs to be elaborated in more details. This aspect is of importance as different Relevance datasets might be created under different judgment conditions. Assessment environment and motivation might affect the results, so that a stable relevance assessment needs to consider global consensus of agreement among judgments.

In the following the basic ideas for considering agreements among multiple judgments will be introduced.

Relying on the above-mentioned issues, this approach adopted the concept of the grade of Agreement from [13].

[6]The implementation details are out scope of this presentation, see voting systems: http://apropat.info/portal/apropat-search-engine/apropat-cognitive-query-model/[7][8]

Definition 3 *(Query Relevance Dataset Agreement & Disagreement)*

Let

- $\langle q\langle \mathrm{D}\rangle\rangle_R$ be a Query Relevance Dataset for the query q as defined in definition (2) with $\langle q\langle \mathrm{D}\rangle_W\rangle_{J_k} = \langle w_1, w_2, ..., w_l\rangle_{J_k}$, $w_i \in [0,1]$, $\forall J_k \in \mathrm{J}$ and $\langle q\langle \mathrm{D}\rangle\rangle = \langle d_1, d_2, ..., d_l\rangle$

- The J disagreement between two assessments in $\langle q\langle \mathrm{D}\rangle\rangle_R$ for some q, is defined in terms of the sum of the absolute differences, and is computed as follows:

$$\mathrm{dist}\left(\langle q\langle \mathrm{D}\rangle_W\rangle_{J_k}, \langle q\langle \mathrm{D}\rangle_W\rangle_{Jy}\right) = \frac{\sum_{i=1}^{l} \left| w_{k_i} - w_{y_i} \right|}{l} \tag{8}$$

- The grade of agreement among the judges in J is defined in terms of the complement of the sum of all pair-wise disagreement within all assessments vectors for the related documents $\langle q\langle \mathrm{D}\rangle\rangle = \langle d_1, d_2, ..., d_l\rangle$:

$$AG\left(\langle q\langle \mathrm{D}\rangle\rangle_R\right) = 1 - \frac{\sum_{i=1}^{m}\left(\sum_{k\neq y} \mathrm{dist}\left(\langle q\langle \mathrm{D}\rangle_W\rangle_{J_k}, \langle q\langle \mathrm{D}\rangle_W\rangle_{Jy}\right)\right) / m - 1}{m} \tag{9}$$

For Example, the agreement among judges involved in assessing the one term query-topic $\langle Cells\rangle$ in context of the document d=ar0001-27-3 in Equation 3 data:

$$\langle Cells\langle d\rangle_W\rangle_J = \langle 0.75, 0.5, 0, 0.25, 0.75, 0, 1, 0, 0.25, 0.25, 0.25, 1, 0.75, 0.75, 0.25, 0.25\rangle$$

$$AG\langle Cells\langle d\rangle_W\rangle_J = 0.591 \tag{10}$$

However, the agreement on this query for of all related Documents D :

$$AG\langle Cells\langle \mathrm{D}\rangle_W\rangle_J = 0.61 \tag{11}$$

In general topics with low, medium or high agreements should be evaluated in their context, when applying them to measure a system performance. However, the agreement with low agreements values might be subject of reformulation relying on the QAV; i.e. Query created Associative Vocabulary of Relevance, which is created by gathering the intuitive and document related associative vocabularies of the query. See Table 1 the first topic; $\langle Cells\rangle$ as an example in the Appendix.

4 Experimental Results and Evaluation

A prototype of the proposed model was implemented as depicted in Figure 2. Implementation details are beyond scope of this presentation. As initial data-source, the ClueWeb2009 containing 29 Million Webpages [8] was used as source for extracting topics related Documents Dataset. Furthermore, LUCENE and APRoPAT Search Engines[7] were also employed in the indexing process, whereas at least 30 documents were extracted for each query-topic. 110 Queries were created based on the following criteria:

- 27 Query-Topics of Type I were created relying on the most frequent 1000 terms in the ClueWeb.
- 23 Query-Topics of Type II were manually constructed relying also on the most frequent 1000 terms in ClueWeb.
- 60 Query-Topics of Type III. 19 queries were selected from TREC-09 and translated manually. The rest (41) were also created by selecting most frequent word randomly.
- All queries were also refined and tested by Google Search Engine to ensure their meaningfulness and validity.
- 21 assessors of different ages and gender were requested to interact with implemented system at different phases and different dates through the web.
- The experiment has resulted in the initial run relevance Dataset of 20.710 relevance assessor feedback and a Vocabulary Co-Occurrence Matrix of 39607 terms distributed in the intuitive, descriptive and document associative vocabulary. Most of the relevance assessor feedbacks are descriptive relevance generated by humans in context of establishing a relevance relationship between a document and documents.
- An overall relevance assessment of the judges for each query was also computed based on the likelihood principal. A relevance Corpus with around 1100 documents was created with multiple-valued assessments in the scale (*Absolutely Irrelevant, Marginally Relevant, Un-decidable Relevant, Highly Relevant, and Absolutely Relevant*).

To ensure the quality of initial dataset, the agreement and disagreement among the assessors for each topic were computed on two agreement levels:

- Agreement on one document.
- Agreement on multiple documents.

[7] A LUCINE based Indexer using Petra Morph and Al-Khalil Morphological Analyzers: http://apropat.info/portal/apropat-search-engine/ [7], [8], [9]

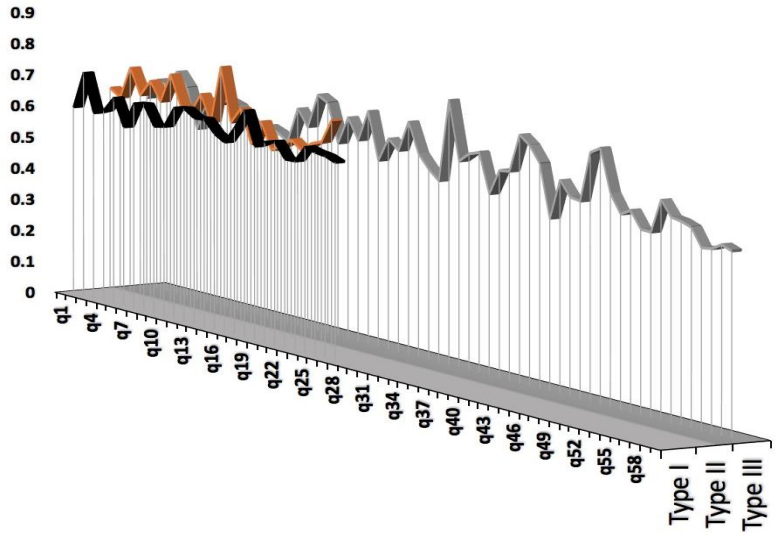

Figure 3
Query-Topics Type I, II and III assessors Relevance agreement on related documents

In the initial run, the grade of agreement depending on the type of the query was ranging from 0.442 to 0.933 on a document agreement level, and from 0.547 to 0.827 on the multiple documents level, provided us with a facility to select a relevance dataset with good agreement in one run. The standard deviation of the assessment depending on the type of considered query indicates a tiny variance, see Figure 3. These results represent stable and useful information for an initial data-source to act as seed for further refinement steps.

However, following some selection criteria such as selecting the queries with high score of agreement would be useful in practical issues in measuring the performance of an IR-System. In this context, it is worthwhile to mention that is it likely to improve all results by considering the other features of Query Associative Vocabulary Dataset at each cycle of refinement; i.e. initial intuitive, descriptive and document associative vocabulary.

Overview and Conclusion

This paper intended to introduce a novel model for query-topic relevance, from assessor and cognitive point of view, in the sense that relevance is a multidimensional cognitive and dynamic conception.

The focus of attention was focused on modeling the concept "Query Associative Vocabulary of Relevance", to stress the value of integrating intuitive, descriptive, multi-valued assessment, and grade of agreement in the process of creating relevance Data. Based on a prototype implementation of this model, a stable query Relevance Dataset was created. Furthermore, as this model differentiates between

different types of topic relevance, it provides a facility of enhancing the quality and augmenting the relevance Data by reevaluating dynamically the resulted Query Associative Vocabulary of Relevance at each cycle of refinement.

Furthermore, categorizing Relevance datasets according to different grades of agreement is important as Relevance Data might give better overview of the performance of considered IR as an inter-cognitive system and the comparison of different Relevance assessment methods in context of consistency and performance is becoming easier.

As human judgments are difficult, time consuming and expensive to obtain; it is important to extract as much advantage from human judgments as possible, and therefore it is planned to increase the machine learning features of this model by enhancing the semi-automatic analysis and query generation aspects of resulted vectors of relevance at each cycle relevance.

In spite of importance of relevance in designing and evaluating Information Retrieval Systems as possible inter-cognitive systems, a consensus on definition is still debatable. However, considering relevance as a multidimensional cognitive and dynamic conception provides researcher with a research track to evaluate the performance of an interactive and inter-cognitive process in terms of the multidimensionality and cognitive aspects of relevance.

References

[1] Mustafa Yaseen, M. Attia, B. Maegaard, K. Choukri, N. Paulsson, S. Haamid, S. Krauwer, C. Bendahman, H. Ferse, M. Rashwan, B. Haddad, C. Mukbel, A. Mouradi, A. Al-Kufaishi, M. Shahin, N. Chenfour, A. Ragheb (2006) Building Annotated Written and Spoken Arabic LR s in NEMLAR Project. Proceedings of LREC, 2006

[2] Azzah Al-Maskari, Mark Sanderson, Paul D. Clough (2008) Relevance Judgments between TREC and Non-TREC assessors. SIGIR 2008, 683-684

[3] P. Baranyi, A. Csapo and Gy. Sallai (2015) Cognitive Infocommunications (CogInfoCom) Springer International Publishing

[4] Kareem Darwish, Douglas W Oard (2003) Probabilistic Structured Query Methods. In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval (2003) 338-344

[5] D. W. Oard, F. C. Gey (2002) The TREC 2002 Arabic/English CLIR Track. TREC 2002

[6] N. El-Khalli, B. Haddad, H. El-Ghalayini (2015) Language Engineering for Creating Relevance Corpus, International Journal of Software Engineering and Its Applications (2015) 9, 107-116

[7] B. Haddad (2018) Cognitively-Motivated Query Abstraction Model based on Root-Pattern Associative Networks, Journal of Intelligent Systems Berlin, Boston, De Gruyter (2018)

[8] B. Haddad A. Awwad, M. Hattab, A. Hattab (2018) Associative Root–Pattern Data and Distribution in Arabic Morphology, Data (2018), 3, 10

[9] B. Haddad (2013) Cognitive Aspects of a Statistical Language Model for Arabic based on Associative Probabilistic Root-PATtern Relations: A-APRoPAT, Infocommunications Journal, Vol. V, 2013

[10] David Hawking (2000) Overview of the TREC-9 Web Track in: in Voorhees, Ellen M., Ed.; Harman, Donna K., Ed. TITLE The Text REtrieval Conference (TREC-9) (9th, Gaithersburg, Maryland, November 13-16, 2000) NIST Special Publication. INSTITUTION National Inst. of Standards and Technology, Gaithersburg, MD. Advanced Research Projects Agency (DOD), Washington

[11] K. Järvelin, and J. Kek¨al¨ainen (2003) IR Evaluation Methods for Retrieving Highly Relevant Documents. Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, 41-48

[12] M. E. Lesk, G. Salton (1968) Relevance Assessments and Retrieval System Evaluation. Information storage and retrieval (1968) 343-359

[13] S. Mizzaro (1998) How Many Relevancies in Information Retrieval? Interacting with Computers (1998) 10, 305-322

[14] S. Mizzaro (1999) Measuring the Agreement among Relevance Judges. MIRA (1999)

[15] Mossaab Bagdouri, Douglas W Oard, Vittorio Castelli (2014) CLIR for Informal Content in Arabic Forum Posts. Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management (2014) 1811-1814

[16] Eero Sormunen (2002) Liberal Relevance Criteria of TREC -: Counting on Negligible Documents? SIGIR 2002 (2002) 324-330

[17] Alvan R. Feinstein and Domenic V. Cicchetti (1990) High Agreement but Low Kappa: I. The Problems of Two Paradoxes. Journal of Clinical Epidemiology (1990)

[18] A. M. Rees and D. G. Schulz (1967) A Field Experimental Approach to the Study of Relevance Assessments in Relation to Document Searching. Cleve- land, OH, Center for Documentation and Communication Research, School of Library Science, Case Western University

[19] Andrew Turpin, Falk Scholer (2006) User Performance versus Precision Measures for Simple Search Tasks. Proceedings of the 29th annual

international ACM SIGIR conference on Research and development in information retrieval (2006) 11-18

[20] E. M. Voorhees (2000) Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness. Information processing and management 36, 697-716

[21] L. Schamber, Linda and M. Eisenberg (1988) Relevance: The Search for Definition. Proceedings of the 51st Annual Meeting of the American Society, for Information Science. 25

[22] L. Schamber (1994) Relevance and Information Behavior, Annual Review of Information Science and Technology, 29, 1994, pp. 33-48

Appendix

Samples of Query-Topics within the Associative Vocabulary (QAV)

The following Table (1) contains some samples of Query-Topics within an Associative Vocabulary of Relevance (QAV) and some extracted values: **Human based assessment**, **Relevance Grades**, and **Assessors Agreement** on certain documents. E.g. based on QAV of the Topic ⟨Cells⟩ represented by assessors feedback, a new query-topic can be proposed such as ⟨Blood Cells⟩ as relevant topic (see Definition 1 and Figure 2). Such query-topics are expected to have higher agreement among the judges; as they have been generated according productive relevance-feedback. On the other hand, Document Associative Vocabulary (DAV) can be utilized to generate documents based relevant queries.

QUERY-TOPIC / QAV-RELEVANCE	DOCUMENT	ASSESSMENTS			
		Human Assessment (non-binary)	Agreement	Relevance Grade	Agreement Category
⟨Cells⟩	ar001-27-3	⟨0.75,0.5,0,0.25,0.75,0,1,0,0.25,0.5,0.25,1,0.75,0.75,0.25,0.25⟩	0.596	0.25	Medium
⟨ Blood Cells⟩		⟨0.75,0,0.75,0.75,0.25,1,1,0.75,0.75 , 0.5,1,0.75,1,0.50 ⟩		0.75	High
⟨ Gas, Prizes ⟩	ar003-57-6	⟨0.25,1,0.75,1,1,0.75,1,0.75,1,0.25, 1,0.75,1,0.75,1⟩	0.823	1	Very High
		⟨⟩			
⟨ Influence of Video Games ⟩	ar000-27-1	⟨0.25,0.5,0,0,1,0.5,1,1,0.5,0,1,0.25, 1,0,0.75,1,0,1,0.5,0.5⟩	0.521	1	Medium
		⟨⟩			