NOTE

Fuhr's Challenge: Conceptual Research, or Bust

David Bodoff
University of Haifa
Graduate School of Management
dbodoff@univ.haifa.ac.il

Abstract

In his 2012 Salton Award acceptance speech, Norbert Fuhr challenged the IR research community to develop IR into an engineering science. In this note, I elaborate on the implications of this challenge, which I interpret as a call for more conceptualization in IR research. The stakes are high, because failing this transition, IR research may never be able to produce generalizable, long-term performance gains, and IR risks losing its significance as a research field.

1 Introduction

In his Salton award acceptance speech at SIGIR 2012, Professor Norbert Fuhr alerted the community to the idea that IR is not developing as an engineering science, and he challenged the community to consider what it would take to change this [1, 2]. He distinguished between theoretical and empirical research, and (constructively) critiqued the state of both. His analysis ended with the following: "*Empirical research is necessary, but it must be accompanied by strong theoretic models. IR evaluations should focus more on answering the why questions, and less on how to achieve good performance for the test collections at hand*" [2 p. 26].

The purpose of this memorandum is to elaborate on Fuhr's central message, which I interpret as arguing for greater conceptualization in IR research. The notion of conceptualization as the basis of scientific research is perhaps unfamiliar in some engineering fields. In this memorandum, I explain its meaning, and elaborate on Fuhr's diagnosis of a problem that, in his assessment, threatens the vitality of IR research as a scientific field, and that is getting worse, not better. I aim to advance the conversation that Fuhr began. I will first present the notion of conceptualization as the foundation of scientific research. Following that, I will discuss implications for IR research, if IR is to make the transition to engineering science.

2 Fuhr's Challenge – The call for Conceptualization in IR research

The main problem Fuhr identifies is the lack of research that identifies variables or features, and the role they play a role in effective retrieval. My recollection is that Professor Fuhr in his oral talk called this "featurization". In the social sciences, it is known as "conceptualization". This process is the basis for all scientific research. Yet IR has not begun to address it in earnest.

For a better understanding of this idea, consider a question that an audience member raised following Fuhr's talk: Did not shipbuilding practice lead shipbuilding theory, with highly sophisticated ships sailing the oceans long before any theoretical understanding? The audience member had understood Fuhr's talk as prioritizing theory over practice (more on this below), and aimed to challenge this position from the shipbuilding example. In fact, Fuhr meant to argue in favor of conceptualization, not in favor of theory over practice, and the challenge from shipbuilding plays right into Fuhr's argument.

Shipbuilding – or, maritime engineering – perfectly exemplifies the conceptualization that distinguishes it from IR and makes it an engineering science. Figures 1a-b show the early-view papers of a leading journal in maritime engineering as of September 2012. Looking at the titles, we can glean a number of characteristics that contrast with what is more common to see in IR papers, and that make maritime engineering more of a science. The first recognizable difference is that each of the maritime papers aims to cogently define and illuminate the workings of a small number of variables. For example, some of the papers shown in the figure regard variables that may improve the strength of a hull – different materials, different kinds of welding joints, different shapes, etc. In maritime engineering, 'hull stress' is a variable that had to be *conceived* or *conceptualized*.

Figure 1a

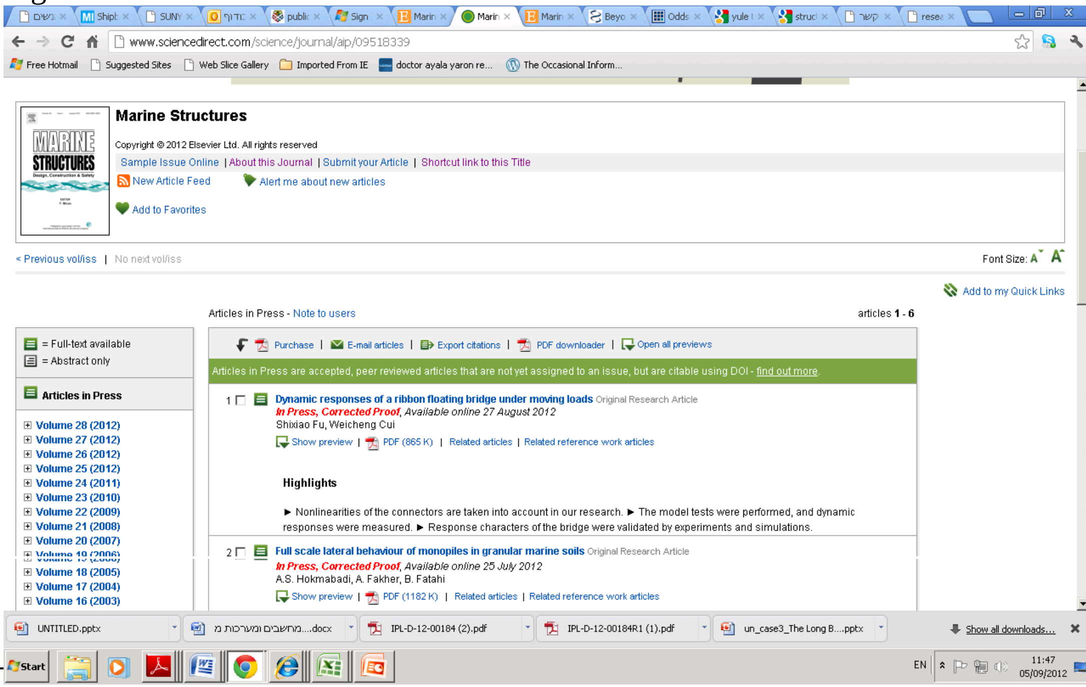
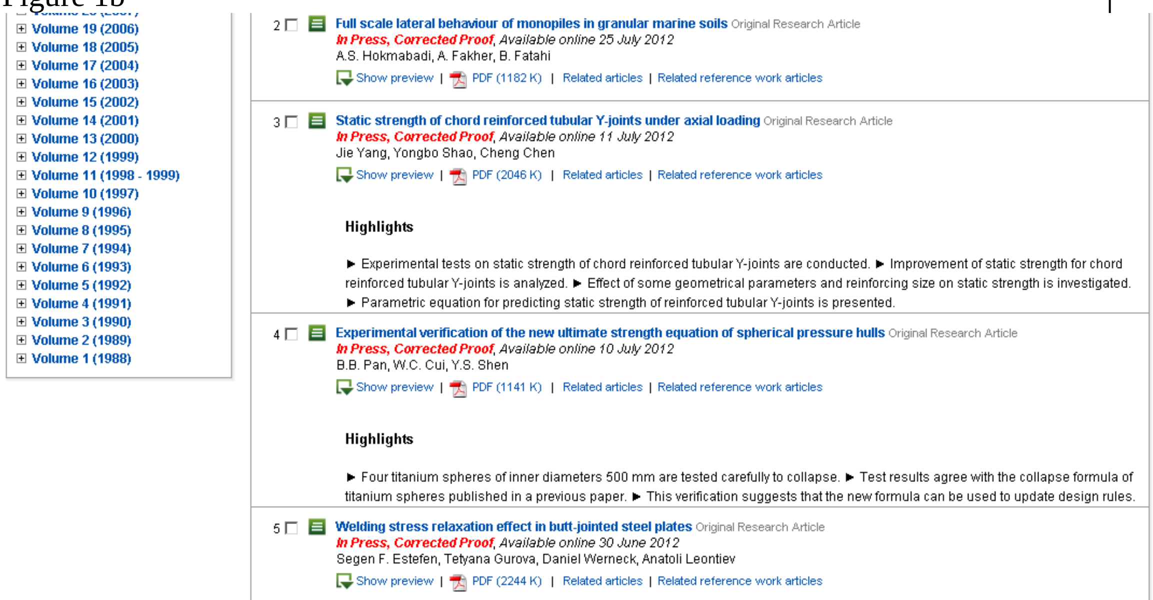


Figure 1b



Conceptualization is the process whereby a variable is defined, both in its own terms, and in terms of how it relates to other variables. Conceptualization is equally theoretical and empirical. A variable's definitions and relations to other variables must be hypothesized, but then those definitions and relationships must also be tested. A definition is not an axiom but is subject to rigorous testing, in the form of tests of content, construct, and predictive validity (or criterion-related) validity [3]. Construct validation includes testing for convergent and divergent validity, which capture the extent to which proposed measures of the proposed variable correlate with one another and not with the measure of other variables. Predictive validity captures the extent to which empirical data supports the hypothesized connections between the proposed variable and other, previously established variables; this, in turn, serves as further evidence that the variable has been meaningfully conceived, i.e. construct validity. All of these empirical tests capture the extent to which the researcher is thinking – modeling the world – in a coherent and useful way. Conceptualization requires researchers to hypothesize and test how a proposed variable relates to specific, previously established variables, often in a causal relationship. This, in turn, requires the researcher to define variables at an appropriate level of abstraction, which allows him/her to relate it to other, previously conceived and established variables. Through this process, using both theoretical ideas and empirical tests, the field develops a "nomological network" of abstractly defined, interlocking variables with defined relationships (hypothetically, one could choose a set of such local models and link them into a single, grand model of everything, e.g. model of ship; but this is not normally profitable, and is never done).

A second characteristic of the maritime papers, and one that follows from the conceptualization process, is that the 'dependent variables' are not *ultimate* ship performance measures (e.g. top speed, miles between refueling, etc.), but simply the neighboring conceptualized variables in the nomological network. In some of the papers in Figure 1, hull strength is a dependent variable, and it is not a measure of ultimate ship performance. It is only one of countless intermediate variables that, through interaction with other variables, may influence a ship's performance on ultimate measures. Some of those other variables may be those affecting the ship's creation of a wake, its response to lateral waves, its creation of drag, and a thousand other variables. No one would expect that every proposed innovation in shipbuilding should be tested on the basis of whether it improves an ultimate ship-level performance measure. The drawbacks of such as requirement are more fully discussed below in section 4.2, but one obvious drawback is that it would require the construction of (a model of) the previous best-performing whole ship as a baseline. This is both impractical and unnecessary for a research paper that explores (say) the effect of construction materials or structures on hull strength.

Conceptualization is at the root of both characteristics outlined above, which differentiate the papers in maritime engineering from those that are more typical in IR. In both IR and maritime engineering, the ultimate aim is good performance. In a conceptual approach, this is accomplished via a detailed and transparent understanding of how a whole 'system' works, i.e. what its moving parts are, and how they inter-relate to produce the macro-level behavior we observe, of the ship, or the retrieval system. This is accomplished in individual steps of research, each of which carefully conceives a small set of locally related variables, and/or tests the validity of such a conceptualization in terms of individual variables and their local relationships. By contrast, without conceptualization, as each individual technique is found to improve ultimate performance relative to a baseline, it is added to an inventory of tecnniques, leaving the problem of which combinations work well. This is where IR has evidently been stuck; I will return to this point in section 4.2.

Returning to shipbuilding, the question was asked, did not sophisticated shipbuilding practice predate theory? One response is that indeed it did, and this practice took the industry pretty far, just as IR systems have come to work pretty well with little conceptual theory development. But then, at some point in the development of shipbuilding research, possibly due to the difficulties in sustaining performance improvements, the conceptualization process was begun; that development would have marked the transition to shipbuilding as an engineering science.

In the case of research in IR, we seem not to have made this transition; this is Fuhr's central point. In fact, when we compare the nature of IR research with the two characteristics observed above in the maritime papers, we appear to have got it fairly backwards: conceptualization is lacking in IR, while testing against ultimate performance measures is *de rigueur*. We turn to each of these in turn.

3 Conceptualization in IR

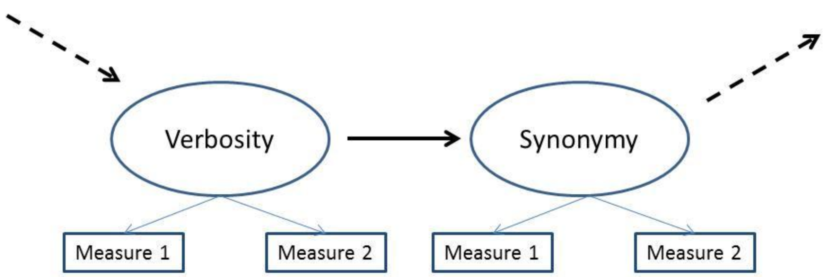
IR research offers little in the way of conceptualized models. In his speech, Fuhr offers a few examples of document 'variables' whose effect on retrieval we would want to understand: 'length, language, domain, genre, structure or the term definition (single words, phrases, named entities etc.)' [1]. Other variables – indeed, wholly other conceptualizations -- may suggest themselves. One might continue the conceptualization process by grouping modeling elements by the problem they address, e.g. elements that deal with the problem of synonymy, elements that deal with the problem that some queries includes terms for only some of the relevant sub-topics, elements dealing with the fact that documents vary in the amount of their content; and so on. This, too, is only a simple beginning example, and only one possibility among many, to give some indication of what real conceptualization might look like in the IR context.

3.1 Example 1: Verbosity

Although I have said – in agreement with Fuhr – that the field suffers from a lack of conceptualization, occasional streams of work do reflect aspects of this process. One example is the 'verbosity hypothesis' versus 'scope hypothesis' of document length [4]. One can imagine elaborations on this idea, such as situational variables that moderate which hypothesis is observed, a variety of remedies or modeling elements that deal with both situations, and so on. One may imagine conceptually-oriented research pursuing the verbosity hypothesis to investigate the relationship between a verbosity variable and a synonymy variable. The researcher may ask: Does verbosity lead to a document's use of multiple synonyms for the same idea (the idea being that people try not to repeat the same exact words)? The answer to this question would ultimately have ramifications for the combination of modeling elements that are effective. For example, if verbosity leads to synonym usage, then for settings/documents in which authors tend to be verbose, the retrieval system need not (ought not) to include synonymy-motivated query expansion. One can imagine theoretical progress along these lines, with various hypothesized variables being related over time. Proposed conceptualizations that are shown empirically to have more predictive connections with related variables will stand the test of time and come to be accepted over others. This is the scientific process.

Continuing the example, research on verbosity would require a way to measure it. This would entail further clarifications, e.g. whether verbosity and scope are two ends of a single construct – in which case the researcher will want to develop a single scale that captures the continuum – or if instead they are conceived as two separate constructs. Either way, the process of defining a measure coincides with further conceptual elaboration. If the researcher were hypothesizing a relationship between verbosity and the extent to which synonyms are found within a document, then a measure of synonymy would also need to be developed. Collecting data on documents' verbosity and/or scope might require judges to manually offer judgements, while the synonymy measure might be more amenable to automatic data collection. Both measures could then be used in a study that tests the hypothesized relationship between them. The empirical results regarding this relationship would constitute further evidence in support (or rejection) of treating "verbosity" as a meaningful variable. The substantive contribution of such work would be insights into the validity of the variables as they are defined, and how they relate to one another. The methodological contribution would be the scales devised for measuring them. This kind of research is often depicted in the form of a "box model", as shown in Figure 2. The ellipses represent abstract concepts, each of which may be measured or "operationalized" in multiple complementary ways.

Figure 2: Example of Conceptual Model



Conceptual research *would* include in-depth analysis of the definitions and relationships among a set of proposed variables, and would most often *not* include tests on ultimate performance measures.

3.2 Example 2: Assumptions Behind Mathematical Models

As a second example of what conceptual research would look like in IR, Fuhr describes "Why-experiments" as concerned with checking a model's assumptions. He gives an example of the binary independence assumption, as a target for possible empirical testing. In such work, the experiment would not be testing an ultimate performance measure, but a measure of how terms are distributed in relevant and non-relevant documents. But his more important example is his reference to Abdulmutalib and Fuhr [5]. This is a more important example because term independence is a *simplifying* assumption that justifies certain calculations, but the broader idea of "testing assumptions" is to test the motivating logic of various mathematical models. The example from Abdulmutalib and Fuhr raises the point that in language models with smoothing, term weight increases with its collection frequency, while in probabilstic models, term weight decreases with its collection frequency, which is related to the probability of occurrence in non-relevant documents. This leads the authors to empirically explore the relationship between collection frequency and relevance, separate from any study of ultimate performance measures. It is worth noting that in other sciences, apparently contradictory results are often an occasion for theoretical refinement that distinguishes two separate and competing effects, as opposed to tests to see which model is right, or a better predictor or performer. The former response to apparent contradictions is common in fields that are steeped in a tradition of conceptual modeling; the latter response is common in fields that are steeped in a tradition of performance improvements, such as IR.

A third example of actual conceptual development – theoretical, this time – is found in

Robertson, van Rijsbergen, and Porter [6], who “consider in some depth the status in the model of the property of relevance and also of the Harter idea of eliteness”. This example revolves around some of the same concepts that are at the core of Abdulmutalib and Fuhr [5]. As noted by Fuhr, these sorts of studies are rare in IR, resulting in a poverty of concepts. The variables that have received some conceptual attention include relevance, eliteness, verbosity, and term occurrence. This is just enough to provide a glimpse of what a conceptual tradition would look like, and how it would ultimately lead to performance improvements.

3.3 Exogenous vs. Design Variables

Until here, we have considered *exogenous* variables that represent characteristics of documents, queries, and words, or more generally, aspects of the environment in which a system will be deployed. A nomological network that consists only of these variables, represents an understanding of how words, relevance, verbosity, eliteness, and so on interact, and would be the domain of information science. Separate from these, are all the *design* variables that represent system features that the system designer controls. For example, synonymy in word usage is an exogenous variable, while thesaurus-based synonym expansion is a system feature; verbosity is an exogenous variable, while document length normalization is a system feature. Our distinction between exogenous and design variables parallels Fuhr’s distinction between controlled and independent variables [2].

4 Points of Difference with Fuhr

In his critique of the state of empirical research in IR, Fuhr notes two trends. First, in his oral presentation Fuhr recalled Armstrong et al.’s [7] result that overall performance of IR systems has not improved over the years [1]. Armstrong et al. attribute this to the practice of comparing against weak baselines, which they -- and Fuhr [1] – view as inappropriate. Second, Fuhr [1, 2] notes a trend of papers in top outlets using proprietary data sets, making it that much more difficult to expect systematic improvements over time. Taken together, he is arguing for tests that use standard collections and compare against previous best-performers. As unobjectionable as these remedies may appear, I believe that they partly miss the mark, by not taking the notion of conceptualized research to its natural conclusion.

4.1 Standardized Testing as Part of the Problem?

There are two aspects to standardized test, standardized test *collections* and standardized *testing*. To the extent that a test collection provides a representative sample of the population of documents and/or queries, it is helpful to the accumulation of scientific results. However, the idea of standardized testing, whereby different techniques are directly compared on a standardized data set, is not common in conceptually driven scientific fields. The simple reason is that by the nature of conceptual research, different researchers study different variables and relationships among them. Our imaginary IR researcher working on verbosity would develop definitions and instruments (i.e. ways of measuring) for that variable. Once these were established,he/she could collect and publicize his/her verbosity data. But only a handful of others will be focused on models that include a verbosity variable. Even those individuals may propose minor adjustments to its definition or to its operationalization, because within the theoretical basis for *their* study, the meaning of verbosity might be subtly different; this would mean that they could not re-use the previous data, or even the previous instrument for measuring it. At some point, there would emerge a consensus about the theoretical definitions and operationalizations of the variable, and this could faciliate the re-use of previously derived verbosity data. Even then, the it would only be useful if the later research could use the same test collection. This is frequently not the case in conceptually driven research, as researchers seek to explore the effects of specific new variables, and not all the newly conceived variables will have been captured in the "standardized" data set. In sum, data re-use is limited -- to other researchers who are interested in the same variable, to research that can be conducted on the same collection, and after consensus had emerged about a variable's definition and operationalization. This is a far cry from the current expectation in IR research, but is typical in the context of conceptually driven science.

The possiblities of data re-use are drastically more limited for researchers studying behavioral variables. In that case, even researchers studying that same variable will usually be unable to re-use the actual data that was collected by previous researchers. For example, suppose one researcher had (defined, devised an instrument to measure, and) collected data on different individuals' query behavior, and a later researcher wished to study the moderating effect of domain expertise on that behavior. The later researcher would necessarily need to collect compleetely new data, i.e. even for the query behavior variable, because both variables must be measured for the same individuals, and the individuals from the previous study are long gone. For this reason, in the social sciences, it is rare to ever be able to re-use actual data. Instead, the holy grail is (merely) re-using others' conceptual definitions and measurement instruments. However, this extreme situation only regards behavioral variables. The situation is better regarding document variables, because documents, unlike human subjects, are still there, and succeeding researchers can measure new variables, as long as the test collection is appropriate.

To summarize, conceptual research leads to convergence in how a variable is to be defined and measured by all researchers in a field, but it limits the extent to which all researchers will use the same data, and the extent to which their research can be directly compared, because different researchers study different variables and relationships. In the prevailing paradigm in IR, researchers offer competing techniques that are tested against one another on identical terms. By contrast, in conceptual work, different researchers offer complementary methods that affect different variables.

4.2 Re-visiting Armstrong et al.: Conceptualization as the Best Hope for Performance Gains

I turn next to the question of comparisons against weak baselines. Fuhr [1] recalls Armstrong et al. [7], who observe that the performance of IR systems does not show any upward trend over the years. Armstrong et al. present this as indicating a lack of any real progress, and attribute it to the practice of comparing new techniques against weak baselines. They suggest – and Fuhr appears to agree -- that papers purporting new, effective methods should compare them against the previous best performers. In our perspective, this recommendation is misguided.

Armstrong et al. themselves raise the possible counter-argument that researchers' job is merely to accumulate an inventory of techniques that work in some sense – e.g. compared to a 'weak' baseline -- with practitioners having the job of finding the best combinations of such techniques. Such a justification would be based (they say) on the assumption that systems that combine various working techniques actually perform better than those that contain just one. Their own simulation shows that this is, indeed, the case, thereby supporting the counter-argument to their central complaint against the practice of comparing to weak baselines. In fact, comparing against best performers may actually hinder progress, because it amounts to a time-dependent greedy search. For example, suppose that query expansion ("A") were proposed earlier than ordered phrases ("B") or term smoothing ("C"), and that neither "B" nor "C" was found to offer marginal improvement over a baseline with just "A". Following their advice, each of the later elements should not be included in an inventory of techniques that are known to work. But this would make it impossible to ever discover that the combination of "B" and "C" offers improvements over a baseline with "A" alone.

The *central finding of fact* in Armstrong et al. is that research has not advanced our understanding of which combinations of techniques work well together. In the perspective advanced here, *this is a classic and predictable indication of a lack of conceptual progress*, and one that is best addressed by conceptual research. When *all* one knows about a technique is that it works in some sense, i.e. compared with a weak baseline, there is no theoretical basis for predicting whether it will have marginal benefit when combined with some other technique. But the best remedy to this is not comparisons against best-performers, but rather, conceptual research. With conceptual research, one learns that a variable "works" by also learning how, i.e. by also learning which other conceptual variables it affects and is affected by. This understanding guides the search for combinations of variables that make sense together. Thus, conceptual research offers the best hope for cumulative performance improvements.

To summarize 4.1 and 4.2, in each individual piece of empirical conceptual research, *one expects to see performance comparisons against weak baselines, possibly using non-standard data*. This sort of work does not tell us who the winner is. Instead, it improves our understanding of how things work. This, in turn, is the best hope for actual, cumulative performance improvements, which requires a way to know (hypothesize) which modeling elements to combine into a full system, and how they logically work together.

4.3 Relationship to Fuhr's "Why" versus "How"

In spite of the above quibbles, I have presented the overarching idea of conceptual research as having been put forward by Fuhr. Because his speech was not explicitly framed using this terminology, it is worth reflecting on where this idea is found in his presentation.

Fuhr's speech promotes "why" research over "how" research. A potential confusion arises in how this relates to Fuhr's other distinction between theory and experiments. In his section 4.3, he writes that experiments "focus on answering how-questions" while "theoretical approaches…deal with the why-questions" [2]. But in his sections 3, 4.1, and 4.2, Fuhr distinguishes between theory and experiments, with the latter being divided into why-experiments and how-experiments. The former quote associates experiments with "how" and theory with "why", whereas the latter quote considers that experiments can be either "how" or "why". I interpret these passages as meaning that *in principle*, there can be both "why" and "how" experiments, but that *in practice,* most experiments in our field tend to be "how" which mostly ignores the conceptual mechanism behind the observed effects. The gist of Fuhr's speech is to promote theory in conjunction with experimental tests of that theory – all of which is aimed at explaining "why". The result of such work is what he calls explanation, validity, and prediction. These are the hallmarks of conceptual research. Conceptually-based results always come together with the logical *explanations* of how they work, which means they always come together with a clear basis for defining the scope of their own *validity*; where their internal logic holds, the results are likely to be valid. When a superficially new task or setting is encountered, the engineer will be able to offer intelligent insights if not exact *prediction*, if he/she can abstract the new situation in terms of variables whose behavior have been previously studied and understood. To summarize, although Fuhr's speeches do not use the same terminology, his main idea coincides with the notion of conceptualization in research, to include both theoretical work that conceptualizes variables and their relationships, and empirical work that tests these hypothesized variables and relationships.

5 Other Trends Away from Conceptual Work

A number of recent phenomena further suggest that the field of IR is shying away from the challenge of turning IR into an engineering science through conceptualized research. First, there is a trend of SIGIR papers that address new phenomena, especially those related to 'social' tools such as Twitter, Facebook, LinkedIn, etc. At present, much of this research is blatantly non-conceptual, in that it defines the phenomena being studied in the concrete terms of particular software tools, such as "factors affecting the likelihood that a tweet will be re-tweeted". When variables remain defined in terms of a particular software, the work is by definition non-conceptual and therefore not a part of science per se. This sort of research may of course yield practical results, and may also set the stage for a later conceptualization of the field. However, there is reason for skepticism; if research on classical IR never makes the transition to an engineering science, and researchers, faced with a lack of progress, choose to 'move on' to new phenomena, there is reason to fear that the transition to science will not materialize with respect to the newer phenomena, either.

Second, an increasing proportion of papers are authored solely by researchers in industry, not academia. I divided SIGIR papers into three categories – all authors are university-affiliated, all are non-university[1] affiliated, and mixed. In 2002, the proportions were 66%, 14%, and 20%, whereas in

[1] Most of the non-university affiliations are commercial, but some are purely research institutes e.g. Chinese Academy of Sciences

2012, the proportions were 47%, 18%, and 35%. Increased academic-industry collaboration is surely a good thing, and research quality is not related to academia versus industry. But the payoffs from conceptual research are (a) long-term, and (b) achieved only through the combined works of a community of researchers. For obvious reasons, such research is historically centered in academia.

Third, there has emerged a separate conference, ICTIR, devoted to IR theory. I view this as a well-meaning response to the desire for scientific progress, but one that reflects a mis-reading of the situation. In my view, IR suffers not from a lack of theory or an over-abundance of empirical work, but a lack of conceptual development. Conceptual development requires both conceptual theory and empirical work to test it. In this logic, to the extent that a separate conference is desirable to allow space for fresh thinking, that separate conference would encourage conceptual research, whether theoretical or empirical.

6 Summary and Conclusions

A convergence of phenomena, some of which were highlighted in Fuhr's 2012 Salton award acceptance speech, indicates a field at a crossroads. A lack of systematic performance improvements begs for a change in course. One possible response is to make systematized performance testing more rigorous, in the hopes of alleviating 'cheating' and achieving the elusive performance gains. A second possible response – already evident -- seeks to diversify into researching newer phenomena, with new – and increasingly proprietary -- data sets. I favor a third response, which is an elaboration of the main point behind Fuhr's speech: the pursuit of conceptual research, which is the basis of scientific progress, and the best hope for long-term performance improvements. Conceptual research requires both theory and empirical work equally. The empirical component is not tests on ultimate performance measures, as has been the predominant paradigm, but of intermediate conceptual structures. This is the nature of scientific work in all fields.

This challenge is independent of the topical agenda put forward by SWIRL 2012 as reported in [8]. That is, the conceptual approach – using both theory and practice – is not limited to classical IR tasks, but is equally applicable to newly framed tasks. One difference between new and old tasks is that, in early stages of IR tasks as in shipbuilding and all constructive endeavors, it may be profitable to begin by 'just doing', i.e. explore the practicalities to learn the issue and frame the problems. Thus, we would expect to see conceptual progress in regard to classical tasks, which are by now well understood, before we would see similar progress in regard to new phenomena such as social, or other topical areas proposed in SWIRL. But at some point, every area of research either progresses down a conceptual path, or else it ceases to be science. IR stands at this crossroads with regard to its more traditional tasks, and this choice will surely be the precedent for how the field matures with respect to today's newly identified tasks.

Fuhr made the case for science. I vote for that.

7 References

1. Fuhr, N., *Salton Award Lecture: Information Retrieval as Engineering Science*, in *SIGIR 2012*. 2012, ACM: Portland, Oregon.
2. Fuhr, N., *Salton Award Lecture: Information Retrieval as Engineering Science.* ACM SIGIR Forum, 2012. 46(2): p. 19-28.
3. Crocker, L. and J. Algina, *Introduction to Classical and Modern Test Theory*. 1986, Orlando: Holt, Rinehart, and Winston.
4. Robertson, S.E. and K. Spark Jones, *Simple, proven approaches to text retrieval*. 1997, University of Cambridge.
5. Abdulmutalib, N. and N. Fuhr, *Language models, smoothing, and idf weighting*, in *Information Retrieval 2010 Workshop at LWA 2010*. 2010: Kassel, Germany.
6. Robertson, S.E., C.J. van Rijsbergen, and M.F. Porter, *Probabilistic Models of Indexing and Searching*, in *Information Retrieval Research*, R.N. Oddy, et al., Editors. 1981, Butterworths: London.
7. Armstrong, T.G., et al., *Improvements that Don't Add Up: Ad-Hoc Retrieval Results Since 1998*, in *CIKM '09*. 2009, ACM: Hong Kong.
8. Allan, J., et al., *Report from SWIRL 2012: The Second Strategic Workshop on Information Retrieval in Lorne.* ACM SIGIR Forum, 2012. 46(1): p. 2-32.