Foundations of Science
https://doi.org/10.1007/s10699-020-09685-x

Why is Information Retrieval a Scientific Discipline?

Robert W. P. Luk[1]

Abstract
It is relatively easy to state that information retrieval (IR) is a scientific discipline but it is rather difficult to understand why it is science because what is science is still under debate in the philosophy of science. To be able to convince others that IR is science, our ability to explain why is crucial. To explain why IR is a scientific discipline, we use a theory and a model of scientific study, which were proposed recently. The explanation involves mapping the knowledge structure of IR to that of well-known scientific disciplines like physics. In addition, the explanation involves identifying the common aim, principles and assumptions in IR and in well-known scientific disciplines like physics, so that they constrain the scientific investigation in IR in a similar way as in physics. Therefore, there are strong similarities in terms of the knowledge structure and the constraints of the scientific investigations between IR and scientific disciplines like physics. Based on such similarities, IR is considered a scientific discipline.

Keywords Science · Information retrieval · Physics · Correspondence · Similarity

1 Introduction

While it may be obvious to researchers (e.g., Van Rijsbergen 1979) in information retrieval (IR) that IR is a scientific discipline, it may not be very easy to explain why it is considered as such to laymen let alone convincing them that IR is science. This is because in the philosophy of science, what science is is a topic of debate (Chalmers 2013). Some philosophers (e.g., Feyerabend 2011) even consider that there is no such thing called science but only specific scientific disciplines like physics, chemistry, etc., as such philosophers consider that there is little commonality between the different scientific disciplines. This situation makes it very difficult for a discipline to claim that it is science since what science is unknown!

Instead of defining science directly, a recent attempt (Luk 2010, 2017) tries to develop a theory and a model of scientific study. This attempt is different from the philosophical approach, which tries to argue what science is. Instead, this attempt treats the definition of science as the construction of a theory and a model, which outline and describe science, by

✉ Robert W. P. Luk
csrluk@comp.polyu.edu.hk

[1] Department of Computing, The Hong Kong Polytechnic University, Kowloon, Hong Kong

Published online: 28 June 2020

identifying the general properties in physics that are applicable to science in general. The theory specifies the aim of scientific study in order to manage the complexity of defining science. To accomplish this aim, scientific study is constrained by a set of principles and assumptions, which are derived from certain part of the aim of scientific study, so that they encourage scientific study to achieve the aim. Apart from these principles and assumptions, the attempt also delineates the knowledge structure of science by a model (Luk 2017) so that different scientific disciplines share the same knowledge structure, thereby supporting the claim that there is such a thing as science, which is regarded as an academic discipline (or an academic subject).

Our novel approach to show that IR is a scientific discipline is to establish that IR is similar to another scientific discipline, specifically physics. This similarity facilitates us to claim that IR is science because the similarity specifies that the aim and the structure of IR scientific study are the same as the aim and the structure of scientific study (Fig. 1) for a well-known scientific discipline, specifically physics, respectively. Why would having the same aim and structure facilitate our claim? This is because the aim and the structure have a scientific character. Specifying the structure of scientific study identifies the components of scientific study, and how they are related. As these components (e.g., scientific knowledge in Fig. 1) are related to science, the structure exhibits a scientific character. The scientific study by itself does not directly try to accomplish the aim of scientific study. Therefore, nine principles, seven assumptions and the knowledge structures are formulated in Luk (2010, 2017) to encourage the scientific study to achieve the aim, which further gives the study a scientific character (as elaborated in Sect. 2.2). To show that IR is a scientific discipline, we need to (1) map the structure of IR scientific study to the structure of a well-known scientific study like physics, (2) show that the aim of scientific study is applicable in the IR context, and (3) show that the principles, assumptions and knowledge structure of scientific study are upheld in IR. Our approach to show that IR is science is reminiscent to the scientific approach in which we find examples as evidence to support our claim that IR is science instead of a logical proof which may still have uncertainties, as the axioms or assumptions deriving the proof may be questioned and as the proof itself may be subject to debate. By collecting more examples or evidence to support our claim, we hope that we are more certain of our belief that IR is a scientific discipline.

Our motivation to show why IR is science is as follows. First, we can convince laymen as well as professionals in other fields that IR is science. Second, we can support the claim that IR researchers are scientists, publishing scientific (journal/conference) papers, attending scientific conferences, participating in scientific societies and carrying out scientific research. Third, we can understand better why IR is science so that this helps us to set our research agenda that makes scientific progress. Fourth, we can understand what the similarities between IR and other scientific disciplines are so that we understand in what sense IR is regarded as a science

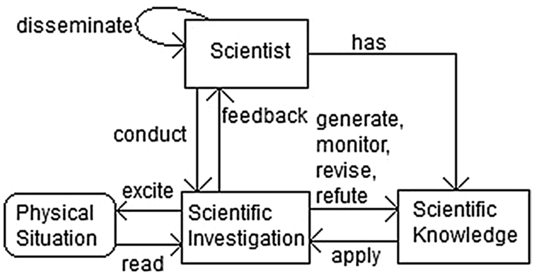

Fig. 1 The basic model of scientific study as a process

thereby helping us to know the nature of IR. Fifth, understanding the similarities and differences between IR and other scientific disciplines also helps us to cross-fertilize ideas between IR and other scientific disciplines. Sixth, the understanding helps the review process to identify the scientific elements of the research papers so that the significance and contribution of these papers are better appreciated, possibly reaching a better review decision about these papers. Seventh, we understand why IR scientific models perform statistically significantly different from random search by random guess, because the scientific models have scientific knowledge that is better than no knowledge (i.e., guessing). Eighth, our understanding shows that theories, models and experiments are all linked up together to form an integral knowledge structure of mature science so that one should not over-emphasize or de-emphasize certain aspect of the scientific knowledge. Finally, this is the first methodology that shows why IR is science, and that can be applied to show why other disciplines are science too. For example, we can use this methodology to help us understand why computer science is science in the future. Therefore, knowing why IR is science is important to many aspects related to IR and science in general.

In this article, we focus on the core part of IR instead of diverting to the human issues related to IR. The human issues of IR are not unimportant, and their studies can be scientifically done (so we are not claiming that they are unscientific). However, we feel that the human issues are not directly relevant to our claim that IR is science, so they are not extensively mentioned here. Some may consider that the human issues may be against our claim that IR is science. For example, users may use the search engines anyway they like, so that it is not realistic to consider that a single evaluation methodology can handle all search situations. However, we assume that the search engine has some intended use(s) and it is evaluated in this respect. We are also not looking at categorical consistency in the evaluation using different users but some percentages of consistency among the users in the evaluation. Since there is risk involved in the IR evaluation, it is typically carried out based on some statistical methodology.

The novelty of this article is the new application of the theory of scientific study by Luk (2017) to establish that IR is a science. The paper by Luk (2017) almost never mentioned anything about IR to establish the theory of scientific study in that paper. Similarly, the paper by Luk (2010) barely mention anything about IR in the context of defining science as a subject. Therefore, this article is completely new compared with the previous two papers by Luk (2010, 2017). More specifically, the aim of IR scientific study is not mentioned, the examples used as evidence to support that IR is a science are absent, the analysis why IR is a science is not carried out in the previous two papers by Luk since those papers were not about IR, and the implication that IR is science is not discussed in the previous papers.

The rest of this paper is organized as follows. Section 2 provides an overview of science, illustrating scientific study using a simplified process model. Based on this process model, Sect. 3 maps IR to science (i.e., to some scientific discipline like physics). Section 4 focuses on answering why IR is science. Section 5 points out some implications that IR is science. Section 6 reviews related work. Finally, Sect. 7 presents the concluding remarks and the future work.

2 An Overview of Science

Science means different things to different people. First, it can refer to a group of subjects under the class, science. Therefore, science is a set of subjects. Subjects in the science set share some commonalities for them to be called science subjects. These commonalities are the properties and the knowledge structures of the subjects. Second, science can refer to the social learning process of generating scientific knowledge. In here, we refer to this social learning process as scientific study. Third, science may refer to the enterprises that organize and build scientific knowledge. Here, we refer to such enterprises as scientific effort. In here, science is referred to as a class of (scientific) subjects which share some commonalities, and which are established by a common scientific study process.

Science as a class of subjects shares some commonalities which are their properties and knowledge structures about experiments, (scientific) models, (scientific) theories and their interrelationships. Apart from these structures and properties, science also shares commonalities in the scientific study process which generates, monitors and applies the scientific knowledge. Such commonalities are formulated as principles and assumptions in Luk (2017), which are mostly linked together by the aim of scientific study in order to achieve such an aim in the long run. When we refer to science as a class of subjects, we are looking at the commonalities not just in the knowledge structure or property but also those in the scientific study process as well because those commonalities ensure that the scientific study generates the scientific knowledge which exhibits the common properties and knowledge structures, shared across different scientific subjects. Since scientific study generates the scientific knowledge, we will first describe a common model of scientific study that is applicable across different scientific subjects, and later we will discuss about the common knowledge structures (Sect. 3.2) and properties (Sect. 3.3).

2.1 Basic Model of Scientific Study

The basic model (Fig. 1) of scientific study [that arises from physics in Luk (2010, 2017)] is that some scientist is conducting the scientific investigation generating the scientific knowledge, which is disseminated by scientists to others for objectivity. Here, we refer to scientific study as the general process of study including the dissemination of scientific knowledge whereas scientific investigation is the specific act of investigating the issues about science without the dissemination of scientific knowledge. Therefore, scientific study is a social learning process, but the scientific investigation can be done without any social interaction.

Figure 1 is a basic model of scientific study, which generalizes the contextual interaction model of Figure 1 in Luk (2017). The scientific study in Figure 1 of Luk (2017), which is reproduced here as Fig. 2 for comparison, is replaced by scientific investigation in order to distinguish scientific investigation from scientific study in general. In addition, Fig. 2 has papers, journals, conferences and proceedings, which are summarized as the directed (dissemination) link from scientists to scientists in Fig. 1 here. The enabling technical knowledge in Fig. 2 is not shown in Fig. 1 here to avoid distraction, but if it is added to Fig. 1 here, it will be attached to the scientist and scientific investigation because the scientist makes use of the enabling technical knowledge to investigate science. In Fig. 2, the physical situation is being measured by the scientific investigation and this corresponds to the link that the (measurement of the) physical situation is being read by the scientific investigation in Fig. 1. In Fig. 2, the scientific investigation may just revise the scientific knowledge implying that the investigation may monitor the scientific knowledge (for revision) and refute the scientific knowledge.

Note that excitation is an optional part of the scientific investigation depending on whether the investigation is a theoretical study. If the investigation is an experimental study, then without excitation we cannot observe the physical situation. For example, placing an

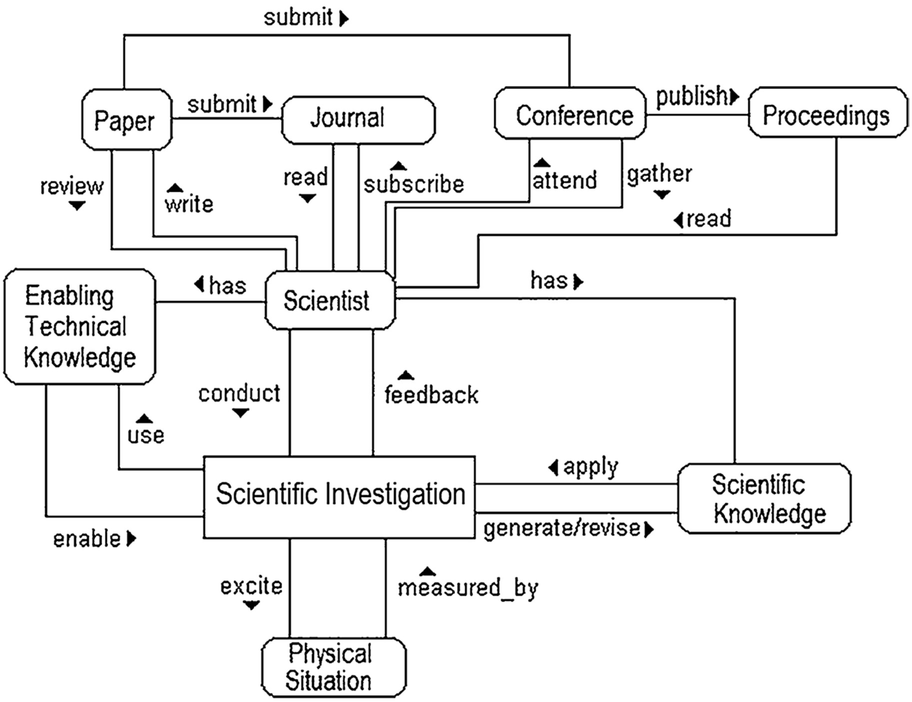

Fig. 2 Contextual interaction model of Figure 1 in Luk (2017) reproduced here for comparison

observer itself may be regarded as an excitation to the physical situation in social science, causing the participants of the experiment to change their behavior. Sometimes the excitation has no bearing on the scientific investigation like using light to read off the meter dials.

In Fig. 1, the scientist conducts the scientific investigation which may generate, monitor, revise and/or refute scientific knowledge. Here, monitor may mean that the scientist tries to validate the scientific knowledge based on an experiment. Some of the experiment may refute or falsify a theory so that the scientific knowledge may need to be revised by the scientific investigation. The data in the scientific investigation may feedback to the scientists who may revise the scientific knowledge (i.e., conducting the [theoretical] scientific investigation) and formulate questions to perform more experiments to probe the physical situation (i.e., conducting the [experimental] scientific investigation).

In Fig. 2, the scientists make use of the enabling technical knowledge in the scientific investigation. For example, scientists use mathematics to describe and quantify measurements of the phenomenon. Another example is that scientists use statistics and probabilities to test the statistical hypothesis in the experiments. Therefore, the mathematics, statistics and probabilities are enabling technical knowledge to help the scientists to formulate and test the scientific knowledge.

2.2 Scientific Character

How can a discipline be considered as scientific? Our novel methodological idea is that the discipline should study like the basic model of scientific study in Fig. 1 since the study will generate, monitor, revise, refute or apply the scientific knowledge shared by the scientific community, just like any other scientific discipline. Since the basic model (Fig. 1) has a scientific character by having certain components like scientific knowledge, scientific investigation and scientists, we need to clarify why they are scientific later, assuming that physical situations are understood by all.

Scientific knowledge consists of theories, models and experiments as described by Luk (2010, 2017) as well as the working scientific knowledge like hypotheses because many scientific disciplines have such types of knowledge. Also, Luk (2017) defines scientists who implicitly or explicitly hold the aim of scientific study as specified in Luk (2017). This aim is very important because it specifies the ideal knowledge that is built by scientific study where the properties of this ideal knowledge can be articulated and given the characteristics that the knowledge is scientific. Since scientists have the long-term aim of scientific study [assumption 4 in Luk (2017)], when they conduct the scientific investigation, the long-term aim will influence how the investigation is carried out, and this gives the investigation a scientific character. For example, one aspect of the ideal knowledge is the quest for reliability. Therefore, the scientists employ methods in the scientific investigation to assess the reliability of the knowledge generated in order to show the reliability of her/his work to the (scientific) community. Because the aim is so important, it is repeated here:

> The aim of scientific study is (i) to produce good quality (measured for example by accuracy, reliability and consistency), objective, general, testable and complete scientific knowledge (as defined in [Luk 2010]) of the chosen domain of study, and (ii) to monitor/apply such knowledge. (Luk 2017)

Fulfilling the aim of scientific study is scientific in the following sense. First, the scientific knowledge should have good quality, which is measured. This is different from everyday knowledge, the quality of which may not be measured. So, the scientific knowledge is concerned with how accurate it is, how reliable it is, etc. so that we can rely on such knowledge to solve problems. Second, the scientific knowledge is objective in two senses. It is objective in the sense that the knowledge is derived from impartial judgment so that we can obtain accurate scientific knowledge (as part of the aim of scientific study). In another sense, the scientific knowledge is accessible by other scientists so that this knowledge is being ascertained by others to reduce our doubt about its objectivity and reliability. Securing objective knowledge is definitely an aim of scientific study. Third, we have to deal with the general knowledge, which has great significance rather than limiting to just the individual facts with limited applicability. Surely, science should be concerned with important knowledge that has widespread significance. Therefore, gaining general scientific knowledge is an aim of scientific study. Fourth, we deal with testable knowledge so that it can potentially be refuted by experiment. However, since experiments established the testable knowledge, we are more certain about the outcome using such knowledge in a testable situation. Finally, the aim of scientific study tries to obtain the complete mastery of the subject even though such an aim may not be achievable in practice. In another sense, the understanding of the phenomenon can be made more complete by scientific study because we try to understand it in detail (measuring the quality of the knowledge obtained) as well as in scope so that we have a more complete picture of the phenomenon as the understanding may link to other fields.

The aim of scientific study is formulated based on past issues discussed in philosophy of science so many of them were debated extensively in philosophy of science. While philosophy usually does not provide any conclusive answers to these issues, they provide insights to many problems in science. Regarding these issues, our aim holds some specific positions and we will make them clear in this article.

First, the scientific knowledge is testable according to our aim. This is concurring the famous view of Karl Popper (1959) on the falsifiability of scientific knowledge. Our perspective is that a scientific theory must be falsifiable although a theory may not necessary be the case when it was formulated.

Second, science looks for general (scientific) knowledge rather than just a set of disconnected facts (Kosso 2007). Therefore, the aim of scientific study seeks general scientific knowledge (e.g., $E=mc^2$) rather than just limiting to specific scientific knowledge which may be scientific facts (e.g., a direct measurement of the speed of light). Having said that, it does not mean that scientific facts are unimportant and does not deserve to be scientifically studied. Instead, the aim of scientific study places some emphasis to look for general scientific knowledge that has widespread significance beyond the specific scientific facts.

Third, the aim of scientific study looks for objective scientific knowledge. The objectiveness (Reiss and Sprenger 2017) of scientific study, scientific knowledge and the reality has been much debated in philosophy (of science). While we cannot claim that the scientific knowledge is absolutely objective, we believe that the scientists strive to make the scientific knowledge as objective as possible through a process that tries to depend less on the subjective value of the individual scientists. This is because objective scientific knowledge is considered to be highly desirable in the aim of scientific study as subjective knowledge may introduce bias causing the knowledge to be inaccurate. Having said that, it does not mean the scientists are not biased but they put effort to make the bias less influential on the subject of study.

Fourth, the aim of scientific study is aimed at good quality of scientific knowledge. Good quality is desired because we want to rely on such knowledge to solve problems. Quality can be measured in terms of accuracy, reliability and consistency. Accuracy can be measured in terms of descriptive accuracy, predictive accuracy, precisions, etc. Therefore, measuring quality can be quite a complicated task. Nevertheless, we do not demand 100% accuracy or near 100% accuracy as in for example scientific realism (van Fraassen 1980) because the reliability may also be an issue where highly accurate scientific knowledge may not be reliable (as in overfitting the data), and because some science subjects may not be able to achieve 100% or near 100% accurate (because of the nature of the physical situation). Instead, we rely on the publication process that tries to find the best accuracy of scientific knowledge that can be attained with acceptable reliability and consistency. As we do not demand the accuracy to be 100% or near 100%, it is important to lower bound the accuracy. Therefore, we formulated a principle that lower bounds the modeling accuracy to be better than random guess so that the scientific knowledge has some utility. As we rely on the scientific knowledge to solve problems, the reliability of scientific knowledge is important. In fact, recently there has been some concerns (Baker 2016) about the reproducibility (i.e., a form of reliability) of the experiments in some scientific studies. Therefore, scientists are concerned about the reliability of their work. In formative scientific studies, many research works only report the accuracy of the results (as in formative IR research). Later, when the field matures, statistical hypothesis testing is usually introduced to show the reliability of the work (as in current research in IR). Therefore, we formulated a guiding principle about reliability in order to encourage scientists to obtain reliable scientific knowledge but at the same time we do not exclude the formative scientific studies to belong to science. It was once thought that scientific knowledge was infallible. However, as philosophy of science progresses, many now hold that scientific knowledge is fallible and therefore it is necessary to assess the reliability of scientific knowledge. Consistency is another important issue that scientific knowledge needs to tackle. In philosophy of science, the Duhem–Quine thesis has been exploiting the consistency of scientific knowledge which is regarded as a system of beliefs. According to this thesis, since scientific knowledge is held as a system of consistent beliefs with a set of assumptions, when an anomaly appears, the scientific knowledge may not be considered to be wrong (because the hard work of the whole knowledge system would be thrown away), but some auxiliary hypothesis is made up to explain the phenomenon based on the current scientific knowledge. The well-known example is the prediction of an unknown planet (at the time) in the solar system instead of considering that Newton's law of gravitation is wrong. Therefore, consistency of scientific knowledge has been an issue in philosophy of science for some time. In the perspective of the aim of scientific study, some theories may be proposed in which there are inconsistencies. While inconsistencies are undesirable, such theories may be the best theories to explain the phenomenon so far (because for example it has achieved very good prediction accuracy). Therefore, inconsistencies are tolerated. However, when we reach the final scientific knowledge about the phenomenon, we do not expect that the scientific knowledge to have any more inconsistencies. If there are, then the scientific knowledge is not final and further work is needed to solve this technical problem. Since many scientific works are still in progress, it may not be unusual to find inconsistencies in some scientific theories. However, we expect that as science advances, these inconsistencies may be resolved in the future. Finally, it is obvious that scientists try to produce a complete mastery of the subject so that the scientific knowledge needs to be complete.

3 Mapping IR to Science

In this section, we apply the aim of scientific study in Luk (2017) to the IR context. Next, we map the IR knowledge structure to a knowledge structure in science (notably physics) to show that they are similar to each other. Physics is chosen not because it is a superior scientific discipline. Instead, it is because physics is an indisputable scientific subject, it is relatively mature, it may be relatively easy to understand, etc. Apart from knowledge structure, we also identify constraints of the IR knowledge, which are similar to the constraints of scientific knowledge in science. Next, we find an IR researcher who is well qualified to be called a scientist according to the definition in Luk (2017). Afterwards, we discuss how IR investigation is similar to scientific investigation. This involves the knowledge about the aim of scientific study as we have specified before. Finally, we discuss how the assumptions about the physical situation are made in science are also made in IR. Effectively, we show how IR scientific study (same as Fig. 1) substantiated with IR works maps to scientific study (Fig. 1), component-by-component. Links in Fig. 1 are the same as those in the IR scientific study because they are general activities (related to science) so that they are applicable to scientific study and IR scientific study. For example, the link that generates scientific knowledge is substantiated by Greiff (1998) in IR. Similarly, the work by Yang and Feng (2016) is an IR example of the monitor link (in Fig. 1), work by Zuo et al. (2012) is an IR example of the revise link, work by Cooper (1995) is an IR example of the refute link and work by Costa and Roda (2011) is an IR example of the apply link. The other links (e.g., excite and feedback) are general activities that are applicable to IR. Thus, the links of IR scientific study are substantiated by IR works, supporting our claim that IR is science.

3.1 Aim

Is the aim of scientific study applicable in the IR context? Yes, this is because the aim of IR scientific study (e.g., in IR papers) can be generalized to gaining good quality, general, objective, testable and complete scientific knowledge about information access (of large repository of documents) as well as monitoring/applying such knowledge. According to this aim, we want to obtain accurate, reliable scientific knowledge (in terms of theories, models and experiments) of information access. Clearly, IR scientific study has been involved in building IR theories and IR models that try to be accurate and reliable, as we will show later (Sect. 3.3). We want to gain general scientific knowledge, for example IR principles have been formulated like the probability ranking principle (PRP) in Robertson (1977). Objective scientific knowledge also is desired, for example confirmation of existing-retrieval-model performance has been studied (e.g., Yang and Feng 2016). Testable scientific knowledge is required, for example PRP (Robertson 1977) is testable. For example, we can ask human beings to assign the probability of relevance and evaluate the rank list sorted by the assigned probability of relevance to see if they produce the best results, assuming that human assigned probabilities are accurate estimates of the probability of relevance. We can also test PRP indirectly by instantiating and deriving retrieval models which are evaluated by experiments. Finally, we desired to obtain the complete scientific knowledge so that we can master the IR subject. Therefore, we can consider that IR scientific study adopts the aim of scientific study in the context of information access (i.e., the domain of study).

3.2 Scientific Knowledge Structure

If IR is a mature scientific discipline, then it should have a knowledge structure similar to that of a mature scientific study (e.g., Newtonian mechanics) in physics for example, according to Luk (2010). That is, a mature scientific discipline should have a theory with principles that are applied to build models, the prediction of which can be measured in experiments. Newtonian mechanics is chosen here to map to IR because it is simple to understand and is a mature science involving theory, models and experiments.

In IR, there is the PRP (as stated by Robertson 1977). This principle can be treated as part of the probability theory of information retrieval (Maron and Kuhns 1960) as in Fig. 3. This principle is applied to rank documents according to the probability of relevance. The probability of relevance is used to derive the TF-IDF term weights in the retrieval models according to Wu et al. (2008) or the language model ranking formula (see Fig. 3). Therefore, PRP is applied to build the retrieval models similar to physics where F = m*a (i.e., one of the laws of motion) is applied to derive the speed of the car on a slanting slope. Similarly for Newtonian mechanics in Fig. 3, we have the laws of motion and the law of universal gravitation (similar to PRP) that are applied to determine the projectile travelling from one place to the other according to the distance formula of the projectile model. The projectile model corresponds to the TF-IDF or Language model, and the distance formula of the projectile model corresponds to the ranking formula based on TF-IDF or the ranking formula of the language model. The frictionless assumption of the projectile model corresponds to the query centric assumption of the TF-IDF and language models as both assumptions are model-specific assumptions.

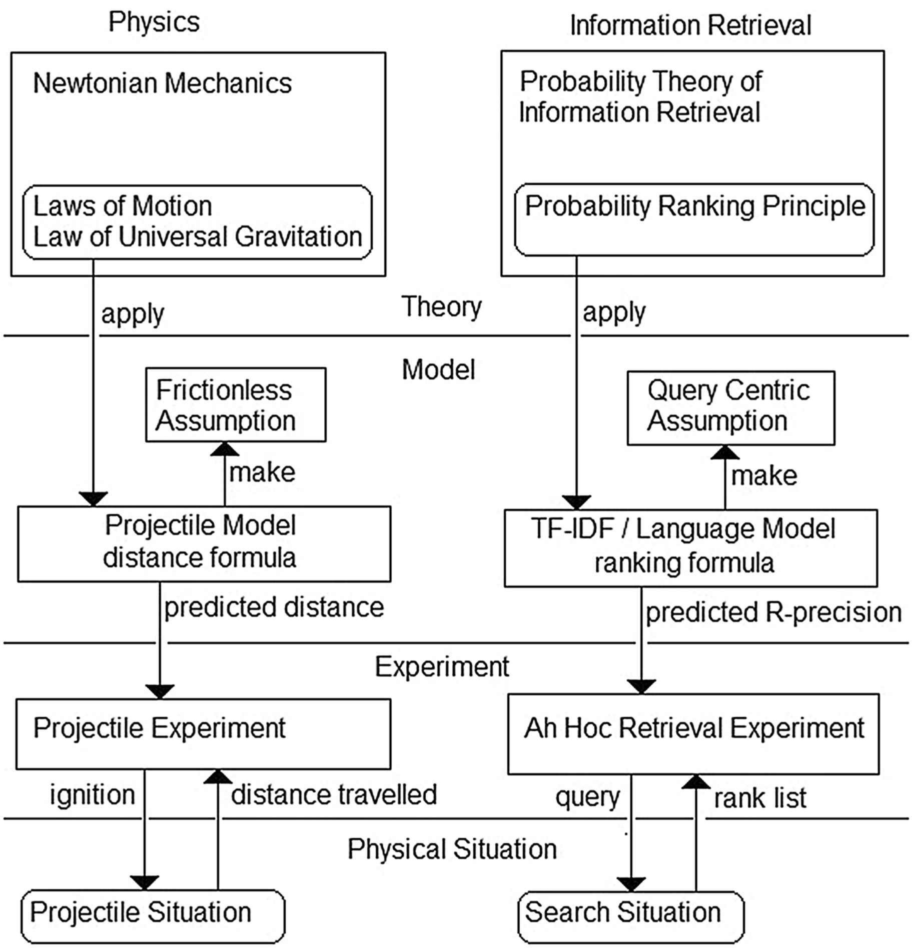

Fig. 3 Parallel knowledge structures between Newtonian mechanics (physics) and the probability theory of information retrieval. See details of the projectile model in Luk (2010)

In IR, the performance of the ranking is measured by experiments as in Wu et al. (2008). Specifically, the performance measurements are precision and recall, which are the measures of the accuracy of the retrieval result based on a set of desirable items. Therefore, the performance measure is a measure of the quality of the IR scientific knowledge, which fulfills part of the aim of scientific study. Thus, in IR, we have the theory linked to the retrieval model by applying the principle (i.e., PRP), and the model predictions are measured by the accuracy of the scientific knowledge as required by science. Hence, this is similar to Newtonian mechanics, and IR has a knowledge structure that is similar to the knowledge structure of a mature science subject (i.e., physics); see Fig. 3 for the parallel knowledge structures.

Note that a principle is different from a physical law. The physical law arises from the abstraction of the observation of data in experiments. The principle is formulated for others to follow. In here, we are saying that the principle corresponds to the law in Fig. 3, where both are used to build mechanical or retrieval models that are tested experimentally.

Therefore, the PRP is similar to the laws of motion in this sense. They are also similar in that they are partly or fully formalized as statements for scientists to apply such laws or principles. Also, we are not claiming that the principles are the same as the laws of motions but that they correspond to each other, just like distance travelled corresponds to the rank list in the experiments in Fig. 3.

In IR, it may first appear that the retrieval models following PRP are not making any predictions (so IR is not a predictive science) but only producing the ranked list. However, according to Robertson (1977) and Dang et al. (2009), if we assume that the probabilities are estimated adequately accurately, then we expect those retrieval models following PRP predict that their ranking produces optimal retrieval with X% performance (say 100%) for MAP, R-precision etc. Therefore, the prediction error of the retrieval models is the difference between X% performance and the actual measured performance of the retrieval models. Since the current retrieval models typically achieve a MAP between 10 and 30% depending on the collection (size), it may appear that the prediction error is quite large. Therefore, there is a need to check whether the prediction performance of the current retrieval models following PRP is statistically significantly better than a random model of retrieval. We will carry out such a test in Sect. 3.3. Note that there are works (e.g., Zamani et al. 2018) that predict the retrieval effectiveness performance typically for subsequent retrieval.

Note that the IR model does not directly predict the accuracy like R-precision in Fig. 3. Similarly, the laws of motions in Newtonian mechanics do not predict the travelling distance accuracy in the projectile model in Fig. 3. The laws of motion only predict the distance travelled and we have to measure the accuracy of the distance travelled. Likewise, IR models produce the ranked lists and the accuracy of the ranked list is measured as R-precision, MAP, etc. Also, note that while the projectile model may appear to be deterministic, IR model actually produces the same result given the same input query. So, the projectile model and the IR model are behaving similarly.

Some may object that we use MAP to measure the accuracy of the search engine because web users typically only want a few highly relevant documents instead of finding all the relevant documents. In this case, a different metric of performance like precision at top 10 documents or the nDCG (Järvelin and Kekäläinen 2002) for the top 10 documents is more appropriate. We note that these performance measures can be used because PRP was shown to be optimal for these measures as well (Dang et al. 2009).

Not all knowledge of IR needs to be put in this form of theory, models and experiments. In fact, many retrieval models do not have any explicit IR theory behind. It may be that these models have not arranged their (scientific) knowledge to show their IR theory, or that these models do not have an IR theory yet. However, there is at least one IR theory in the IR discipline, the knowledge structure of which is similar to physics, a mature scientific discipline, so that we may claim that IR is science. Similarly, not all physical phenomena have developed scientific theories and models in physics, and we do not map IR to those theories or models.

There is no guarantee that the IR model satisfying the mature science structure requirement of having linked to some theory will have the best effectiveness performance. However, the social learning process, IR scientific study, will look for the best performing theory or model. In addition, if such IR models are not providing the best performance, then there is obviously an open research question as to why the IR theory cannot produce the best model. Is there something wrong with the theory? Or, is the estimation not accurate enough? Why are other models capable to produce better results? Do the other models have an implicit or more important, accurate IR theory waiting to be discovered? While the scientific framework cannot provide us with the answer, it can generate many questions for us to consider more deeply, and these may lead to important advancement in the field. In IR, there is the additional sign post by Robertson (1977) and Dang et al. (2009) that if the probabilities are estimated accurately, then PRP specify a ranking that is optimal for many different performance measures (like Mean Average Precision, Precision for the top *n* documents, etc.). Therefore, if the model based on PRP is not the best, then this is a surprising result according to Robertson (1977) and Dang et al. (2009). Could this be due to the estimation? Could it be that there is something wrong with the simplifying assumption so that the estimated probability for ranking is different from the ranking of the probability of relevance? Is there a more powerful theory than the one based on PRP? If so, why would it be more powerful? For example, the generative theory of IR (Lavrenko 2009) is an alternative to PRP. However, the generative theory does not claim to have any principles but only a few hypotheses. Even if we consider these hypotheses as principles for comparison purposes, the generative theory is not related to the performance of the retrieval models, so it is hard to attribute the perform to the principles (or hypotheses). As a result, it is hard to compare theoretically the generative theory of IR with PRP. Empirically, they can be compared by observing that the retrieval model belonging to the theory can perform better than the other model of the other theory. There are other IR theories [e.g., Quantum IR theory (Van Rijsbergen 2006) or Axiomatic Theory of IR, e.g., (Zhai 2011)] available but our task is to pick one particular theory that can illustrate that IR can be mapped to science instead of exhaustively going through every IR theory as different IR theories are at different stages of development. Unlike PRP, these other IR theories do not in general make predictions about retrieval effectiveness so that we cannot claim our models to be performing as predicted with a certain amount of error. Therefore, we did not correspond Newtonian mechanics with these IR theories.

There are also assumptions in IR theory just like those in scientific theories or models. For probabilistic IR theory, Kolmogoroff axioms may be considered as the basic assumptions. However, quantum IR theory may not necessarily assume them as such theories or models may violate Bell's inequality. Another common but lesser known assumption in IR is the query centric assumption by Wu et al. (2008). This assumption is not true all the time and it only applies to the relevant documents about 80% of the time. However, this assumption makes the IR modeling more tractable. This is an example of another type of assumptions (called model-specific assumptions) that appear in scientific modeling, which also appears in IR. Overall, IR theories and models make assumptions similar to scientific theories and models.

Apart from principles and assumptions, IR also has definitions as in science. Notably and recently, Zobel (2017) was concerned that past descriptions of IR are too restrictive. He offered a definition that suggests IR as "a study of techniques for supporting human cognition with documents, using material that is sourced from large document collections" (Zobel 2017). This definition is much broader than previous definitions or descriptions that mostly purport IR as finding documents based on some information need. One of the concerns of writing definition is whether the definition over-generalizes or over-specializes. While focusing on finding documents may appear over-specialized, it is possible that (a) the study of document clustering can be considered as preprocessing in support of browsing, which is a form of information access, and (b) the study of link analysis is to find reliable documents for retrieval supporting human consumption. Therefore, it is not clear whether past definitions of IR are really over-specialized. Further study is needed to come up with an acceptable definition of IR that neither over-specializes nor over-generalizes

IR (e.g., document processing for human cognition). This is not going to be easy as the IR field evolves, and many definitions can be proclaimed.

3.3 Scientific Knowledge

Apart from the knowledge structure, some common principles were formulated by Luk (2017), and they specify some properties of the scientific knowledge. If IR is a scientific discipline, then IR knowledge should obey such common principles as shown in Table 1. These principles were formulated to encourage the scientific investigation and scientific knowledge to achieve the aim of scientific study. To show the relationship between the specific principle to the specific aim of scientific study, a column on the related attributes in the aim of scientific study is added in Table 1. As discussed in Luk (2017), no principle is formulated for the completeness attribute of the aim of scientific study because it was thought to be obvious. Note that the quality of scientific knowledge is measured in terms of accuracy, reliability and consistency. Therefore, several principles were formulated for the quality attribute of the aim of scientific study in Table 1.

The first common principle [principle 1] is the basic principle of generalization (Table 1). This principle in scientific study requires the theory generalizes the applied models, which generalize the corresponding physical situations of the experiments. Because of this principle, the PRP needs to be a generalization of more than one retrieval model. Actually, the PRP is applied to derive several versions of the TF-IDF term weights, including the BM25 based on a model of relevance decision making (Wu et al. 2008). Later, PRP can be shown to derive the language model (LM). This is done first by showing that the query likelihood can be derived from the log-odds ratio after making two simplifying assumptions, based on the work of Lafferty and Zhai (2001) and the work by Azzopardi and Roelleke (2007), or the work by Luk (2008). After showing that, the log-odds ratio can be shown to be rank equivalent to the probability of relevance (which is specified in the PRP), as follows:

$$p(r|q,d)/p(\bar{r}|q,d) = 1/p(\bar{r}|q,d) - 1 \propto 1/p(\bar{r}|q,d) \propto -p(\bar{r}|q,d) \propto p(r|q,d),$$

where r is the relevance value, $\bar{r}$ is the non-relevance value, d is the document, q is the query, $\propto$ is the rank equivalence relation and $p(r|q,d) + p(\bar{r}|q,d) = 1$. Therefore, PRP is a general principle applied to more than one (successful) model. The TF-IDF term weights, BM25 and LM have been demonstrated before successfully as state-of-the-art retrieval models for more than one test collection. Therefore, these models generalize more than one physical situation in more than one experiment. Thus, the basic principle of generalization holds.

The second common principle [principle 2] is the basic principle of modeling accuracy. This principle specifies that a scientific model should perform better than by random guess. In IR, this is rarely shown explicitly. In this article, we consider how we perform a random guess given a query similar to a realistic search situation. Specifically, we consider gathering a sample of documents that contain at least a query term from the collection. Then, we perform random sampling of say 1000 documents from this sample and measure the retrieval effectiveness as the performance of a random model guessing the retrieval result. Thus, this would be a more realistic comparison of whether the existing model performs better than direct random sampling documents from the collection given that the random guess made use of the query.

Table 1 Summary of the principles of scientific study in Luk (2017)

No.	Name of principle	Related attributes in the aim of scientific study	Content of principle
1	Generalization	General	The theory generalizes the applied (related) models which generalize the corresponding physical situations of the experiments
2	Modeling accuracy	Quality	A scientific model should achieve statistically significantly better prediction accuracy than random guesses using the appropriate minimal prior knowledge
3	Empiricism	Testable	A scientific theory must be directly and indirectly based on evidence from experiments, which supports or potentially falsifies the theory
4	Theoretical objectivity	Objective	A scientific theory and its supported scientific models must be explicit so that they can be communicated to other scientists unambiguously and should be fully or partly formalized for reasoning and testing inconsistencies
5	Theoretical consistency	Quality	A scientific theory and its supported scientific models must not be inconsistent with each other and with themselves
6	Immutable laws and principles	Quality	Principles and (physical) laws in (scientific) theories should not change in time
7	Objective experiment	Quality and Objective	An experiment should not be intentionally biased to obtain a particular, favored outcome by manipulating the experiment
8	Reliability	Quality	Scientists should use methods to assess the reliability of their (working) scientific knowledge obtained by conducting scientific studies
9	Investigation objectivity	Objective	Scientists should enable other scientists to carry out the scientific studies for independent verification

We have implemented such a random search model (i.e., random guess) in our retrieval system. To test that it performs worse than the common IR model (e.g., BM25), we run the test for the WT10g, GOV2 and Clueweb09 test collections. The WT10g contains about 1.6 million web pages, the GOV2 has about 25 million web pages and Clueweb09 (Category B) has about 50 million web pages. We used 50 topics in GOV2 (terabyte 2006) and 50 topics in Clueweb09 (Web track 12) as the training data for the BM25 retrieval model. We estimated the parameter values for the BM25 model using terabyte 2006 data by performing a grid search for the best parameter values, and we apply this model using the same parameter values to the other topics of GOV2 and the topics in WT10g. Similarly, the other topics in Clueweb09 are used to compare the performance of BM25 (a common retrieval model) with the random search model guessing the retrieval based on the query information.

Table 2 shows the MAP performance of the BM25 model and the random search model for three document collections. The MAP is measured based on the test topics only. The MAP of the BM25 model is higher than the corresponding MAP of the random search model and the MAP differences are statistically significant at the 95% confidence interval for all three collections. Therefore, we conclude that the BM25 model, representing our scientific IR knowledge, is better than the random search model. To be more precise, the BM25 model makes less prediction error than the random search model. This is because the predicted MAP performance is say X% and the actual MAP of BM25 is higher than the random search model, so that the prediction error, which is X% minus the actual MAP, is smaller for the BM25 model compared with the random search model. Therefore, we can conclude that the BM25 model, as a form of scientific knowledge, is more accurate than the random search model, and this experiment has evidence to support that IR knowledge satisfies the basic principle of modeling accuracy in Luk (2017).

There are other common principles related to the scientific knowledge and they seem evident that they are satisfied in IR. First, the basic principle of empiricism [principle 3] requires IR theory to be testable. Clearly, there is no guarantee that PRP leads to the top retrieval model, so PRP needs to be tested. Second, the basic principle of theoretical objectivity [principle 4] requires the scientific knowledge to be partly formalized so that it is communicated to other scientists for reasoning and testing inconsistencies. Indeed, PRP was published in a paper and it was partly formalized. Third, the basic principle of theoretical consistency [principle 5] requires the IR theory to be consistent with the supported retrieval models. In IR, actually BM25 and some TF-IDF term weightings were derived and instantiated from PRP by Wu et al. (2008) after PRP was formulated for over three decades or so. Therefore, the supported retrieval models are obviously consistent with the probability theory of IR. Finally, the principles in IR are meant to be immutable and do not change in time [principle 6] unlike some legal principles (so that science cannot be

Table 2 Mean average precision (MAP) results of the BM25 retrieval model and the random model for the test topics, trained by the (other) training topics

Collections	WT10g	GOV2	Clueweb09
Training topics	Same as GOV2	801–850	151–200
Test topics	451–550	701–800	1–150
BM25	.2084*	.2968*	.0969*
Random	.0443	.0421	.0016

Note that * indicates that the MAP is statistically significantly different from the corresponding MAP of the random search model with a 95% confidence interval claimed by having principles alone). For instance, the probability ranking principle (Robertson 1977) was formulated over four decades ago and it has not been changed (when the independent relevance assumption holds) for the basic ad hoc retrieval settings. However, it has been modified for interactive retrieval (Fuhr 2008), and it is found to be non-optimal for adversarial search settings (Basat et al. 2015).

3.4 Scientists

According to Luk (2017), scientists are those who are:

(a) capable to acquire (working) scientific knowledge of the domain; and
(b) capable to acquire the enabling technical knowledge for her/him to conduct scientific study; and
(c) use methods and/or methodologies that can accomplish some or all aspects of the aim of scientific study.

According to these requirements, most IR researchers are scientists because they have the relevant background to perform all three requirements above. For example, Robertson has a background in mathematics. He then studied IR with Karen Spärck Jones who has already worked on IR. So, it is not difficult to see that Robertson at the time is capable to acquire the (working) scientific knowledge. Second, he is capable to acquire the enabling technical knowledge, as he was a mathematician who probably has some training in probability and statistics. Third, would he apply methods and/or methodologies that can accomplish some or all aspects of the aim of scientific study? In the past, he worked on PRP, so this is an indication that he seeks general scientific knowledge fulfilling part of the aim of scientific study. He also developed the retrieval model based on BM25 term weighting. It was shown that BM25 is one of the most successful retrieval models in ad hoc retrieval. So, he tries to acquire accurate knowledge in IR. He has been working on IR evaluation commenting on the reliability of the evaluation (e.g., GMAP by Robertson 2006), so this is again a sign showing that he tries to use methods and/or methodologies to accomplish some aspect of scientific study. He disseminates his research by publication and by participating in TREC so that people can reproduce his work for objectivity. He also worked on testable knowledge like PRP, which may turn out to produce not-effective retrieval models. Apart from completeness which is difficult to achieve for many fields (not just IR), Robertson has tried to use methods/methodologies that can accomplish almost every aspect of the aim of scientific study (apart from completeness). Therefore, Robertson can safely be considered to be a scientist by the definition in Luk (2017). In general, the scientist definition is more relaxed to recognize a scientist as it only requires some aspects of the aim of scientific study to be fulfilled. For example, Salton, Spärck-Jones, Croft, Van Rijsbergen, Fuhr and Lafferty are all scientists according to the definition.

3.5 Scientific Investigation

There are different types of scientific investigations. One type is theoretical study which involves the theory and the model but without any experiment. In physics, Einstein's papers on special/general relativity are examples of theoretical studies. IR scientific study has theoretical studies too, e.g., Robertson (1977), Dang et al. (2009), Lafferty and Zhai (2001), Azzopardi and Roelleke (2007) and Luk (2008).

Apart from theoretical studies, we also have investigations that involve experiments. There are different subtypes of such investigations. First, experimental studies (e.g., Huston and Croft 2014) may verify the retrieval models, which is very common in IR. Second, experimental studies (e.g., Greiff 1998) may try to test or construct the theory. Third, experimental studies (e.g., Spärck-Jones 1972) may just involve the experiments without theory or model. For all these subtypes, experiment is an essential part of the investigation. In this regard, the scientific investigation may be considered to follow the scientific method (SM), so that work by Luk (2010, 2017) may be considered to encapsulate SM. Since IR is empirical and we have shown how IR makes predictions in Sect. 3.2, it is not difficult to see that IR scientific investigations can follow the SM. However, one important difference is that Luk (2010, 2017) stresses the general notion of reliability of the investigation rather than the more restricted form of reliability (i.e., reproducibility) as in SM so that historical science can be included in Luk (2010, 2017).

In Luk (2017), there are some principles that govern how the experiments should be conducted and IR needs to obey these principles. First, the basic principle of objective experiment [principle 7] requires the experiment to be done without any bias to favour any particular theory or model over others. Such principle should be upheld in IR because it is about the fairness of the experiment. For example, Lin (2018) is concerned with using weak baselines to favour the proposed retrieval model to demonstrate better performance. It represents work that is concerned with the fairness and impartiality of the experiments. Therefore, it is an example of upholding the principle of objective experiment in Luk (2017).

Second, there is the guiding principle of reliability [principle 8], which specifies that methods are used to assess the reliability of the work. In IR papers, frequently statistical tests (e.g., Zhai and Lafferty 2004) are done to show the statistical significance of the work. These tests indicate the reliability of the claim that the performance of one retrieval model is different from another retrieval model. Apart from reporting statistical significance results in IR papers, some papers (e.g., Yang and Feng 2016) report the reproducibility of the results, which is a form of reliability. Therefore, we believe that the guiding principle of reliability is upheld in IR.

Third, there is the guiding principle of investigation objectivity [principle 9]. In our experience, we have asked other individual IR researchers about their implementation of retrieval models and all of them have replied how the implementation was done. Also, for reproducibility, some IR researchers (e.g., see (Croft et al. 2010) for Galago) provide their software (like Indri 2013 or Terrier 2019) for others to verify its performance. Therefore, we believe that the guiding principle of investigation objectivity is upheld in IR.

Fourth, assumption 3 (see Table 3 for a list of assumptions) in Luk (2017) demands the researchers to strive to make unbiased, adequately accurate observations in experiments. For most of the experiments involving retrieval model verifications, there does not seem to be any problem with satisfying this assumption as the performance is read off mechanically. However, there is concern (e.g., Fuhr 2017) that some IR papers report results with too much precision that can be supported.

Finally, assumptions 1, 2 and 4 in Luk (2017) are usually satisfied. Assumption 1 requires the scientists to be adequately trained. For instance, Robertson being cited as an example scientist is well trained in IR. Assumption 2 requires the scientists to communicate accurately. This is intended to be true by the scientists although from time to time unintended inaccurate communications in papers may be found. Further publications may be required to clarify the communication so that assumption 2 is salvaged. Assumption 4 is

Table 3 Assumptions in the theory of scientific study by Luk (2017)

No.	Name of assumption	Related attributes of the aim of scientific study	Content of assumption
1	Sufficiently trained	N.A.	Scientists are sufficiently trained to conduct or to be enabled to conduct scientific studies
2	Accurate communication	Quality and objectivity	Scientists express their work accurately in scientific communication
3	Unbiased, accurate observation	Quality	Scientists strive to make unbiased, (adequately) accurate observations in experiments
4	Adoption of the aim of scientific study	All attributes	The domain of study using scientific studies adopts the aim of scientific study
5	Causality of phenomenon	N.A.	In a scientific study, the phenomena observed in the physical situation have causes
6	Explanatory power	N.A.	A phenomenon in a physical situation can be explained by some theory or model
7	No magic	N.A.	If similar or identical physical situations occur, then similar or identical situations will produce similar or identical (probabilistic) distributions of outcome, respectively

Note that N.A. stands for not applicable about upholding the aim of scientific study in Luk (2017) for the domain of study, which is obviously needed if the study has a scientific character (Sect. 2.2).

3.6 Physical Situation

The physical situation in IR scientific study commonly makes assumptions 5, 6 and 7 in Luk (2017). First, assumption 5 specifies that there is some cause of the phenomenon observed. In IR, many events are assumed to be causal. For example, a document is relevant to a topic because the document has information related to the topic. The reasoning is that there is information (i.e., the cause) in the relevant document, causing the relevance judgment to signal that the document is relevant (i.e., the phenomenon). Therefore, by measuring the related information, we may be able to predict which document is relevant or not to a topic. Such a causal argument of relevance judgment is implicitly used in many retrieval models. For example, the TF-IDF is a measure of the information related to the topic. The inverse document frequency (IDF) measures the specificity of the term, which if it is very discriminating, implies that the occurrence of such a term implies that the topic is related somehow. The term frequency (TF) factor is a measure of how strong the signal is carried by the term. If the term occurs many times, then the likelihood of at least one occurrence of the term that is relevant is higher, causing the overall relevance judgment to be "relevant". Thus, TF is a positively related signal for relevance judgment. Overall, IR research works do make assumption 5.

Assumption 6 assumes that there is a theory or a model to explain the phenomenon. In IR, researchers in general make such an assumption if the phenomenon is important and interesting. While some phenomenon may be not explained yet by a theory and a model, it is believed by IR researchers that such phenomenon can be explained later. For example, relevance is a very elusive concept to define and capture, but this notion as a phenomenon is still explored by researchers like Saracevic (1975). Further studies (e.g., Wong et al. 2001) try to tackle one aspect of relevance based on what the common meaning is for "aboutness". Therefore, even though it is very difficult to grasp the common concept of relevance, IR researchers still study and try to come up with some theory or model of the phenomenon (i.e., what is the common notion of the concept, relevance).

Assumption 7 assumes that similar or identical situation may produce similar or identical distributions of outcome in the situations. IR researchers commonly make such an assumption so that they can reproduce their experiments. For example, it is commonly assumed that evaluators with extensive knowledge background of the topic make similar relevance judgment as other knowledgeable evaluators despite the fact that agreement of relevance judgment between users is known to be less than 100% (e.g., Al-Maskari et al. 2008; Damessie et al. 2017). Otherwise, we may need to have relevance judgments from more than one evaluator for each judged document, making the evaluation process very labour intensive. Note that for some corpora, multiple evaluators are indeed used to build the test collection (e.g., CF corpus). However, this is too costly for large collections and most large test collections assume different evaluators make similar relevance judgments. Note that we do not need the relevance judgments to be categorically identical between different users. As long as the consistency is well above the best performance achieved by current search engines, we know that there is still a wide margin that the search engines need to be improved before consistency of relevance judgment becomes an issue in the evaluation of search engines. At present, the agreement of users in relevance judgment is about 60–70% (e.g., Al-Maskari et al. 2008) which is much higher than the best MAP performance of search engines (typically 30+% for title queries using fully automatic retrieval without any training). Therefore, the consistency of relevance judgment is not a significant issue at present, especially for large document collections as the best MAP performance tends to be lower because the retrieval tasks become more difficult to perform as there are more noise (in the documents) that the search engine finds it hard to differentiate from the signal.

4 Putting the Claims Together: Why?

After mapping IR to science, we are in a position to answer the question: why is IR science? First, the IR scientists uphold the aim of scientific study when they carry out their scientific investigations similar to other scientists upholding the same aim. Why is upholding that such an aim can claim the discipline as scientific? This is because the aim has a scientific character as explained in Sect. 2.2. Note that the aim requires that knowledge has to be scientific, which means that the knowledge is related to theories, models, experiments and physical situations.

Second, the IR scientific investigations are also scientific because the scientists by definition will use method or methodologies (in the scientific investigations) to accomplish some or all of the aim of scientific study. Since the aim of scientific study has a science character as stated in Sect. 2.2, this in turn gives the investigations a scientific character. For example, since the aim of scientific study is to produce objective (scientific) knowledge, the scientific investigation needs to be disclosed to others so that this fulfills the guiding principle of investigation objectivity according to Luk (2017). One may wonder whether there are such scientists that may use the aim of scientific study to drive the scientific investigation in IR. Therefore, we have cited Robertson as an example IR scientist who is shown to investigate IR, fulfilling most parts of the aim of scientific study.

Third, the scientific investigations generate IR scientific knowledge. The structure of such scientific knowledge is similar to the structure of scientific knowledge in physics, which is considered to be a science subject. Why would having a knowledge structure similar to physics (a science subject) help us to claim that IR is science? This is because this is the commonality between different science subjects. Without this commonality, science subjects may not share any common characteristics, in which case there may not be science at all. Also, such knowledge structure also fulfills part of the aim of scientific study so that the scientific knowledge is organized from the most specific (in the experiment) to the most general (in the theory).

Fourth, apart from knowledge structure, the retrieval models in IR perform better than random guess, which is required by science. Random guess produces a lower bound performance for the scientific model to overcome. It is because the scientific model has some scientific knowledge that is better than no knowledge represented by random guess. This goes back to the aim of scientific study, which tries to gain quality knowledge that we expect to be better than no knowledge.

Finally, why can we claim IR is science after establishing that IR scientific study is similar to scientific study (Fig. 1) say in physics? This is because the aim will constrain IR scientific study to produce IR scientific knowledge (which consists of theories, models and experiments) where such knowledge has a similar structure as scientific knowledge in a scientific discipline (i.e., physics). As a result, IR is science (as a scientific subject or as a scientific discipline).

5 Implications

Currently, IR is considered as a sub-discipline of computer science, so this supports the claim that computer science is science (Denning 2005). However, it would be quite costly to show that every sub-discipline in computer science is science before computer science is claimed to be a science. Therefore, we want a more efficient way to show that computer science is a science. For instance, to claim computer science is science may only involve claiming that the core sub-disciplines of computer science are science as the core sub-disciplines are applied in every aspect of computing. As this topic is involved, we leave it for future work.

One implication of this work is to encourage IR researchers to build a more complete scientific discipline than the current one. For example, is there an overarching principle that can be applied to build divergent sets of retrieval models including those that are not probabilistic ones [e.g., pivoted document length normalization (Singhal et al. 1996) or MATF (Paik 2013)]. Another example is to determine the upper bound performance limit of retrieval models with the lower bound performance being set by the random model. How can the upper bound performance be set? Is the upper bound performance of the model limited by how human relevance judgment is made (e.g., Al-Maskari et al. 2008)? If so, can a group of human judges be used to estimate the upper bound performance of retrieval models? From these examples, this work opens many issues for IR researchers to investigate that can make the field more completely scientific.

Another implication of this work is on the review process of IR. While it is very desirable to have all the components of a mature science in a single paper, it is very difficult to be that inclusive. It is also not very practical to require papers to achieve all aspects of the aim of scientific study by Luk (2017) because the aim is supposed to be a long-term aim that may not be attainable (although there are methods that direct towards achieving such an aim). We believe that the review process should recognize the significance/contribution of the paper reaching some aspects of the aim of scientific study or some part of a mature science, so that the scientific knowledge is established over time. For instance, special relativity was published as a paper without any experimental support, but it was allowed to be published because its significance/contribution is recognized. Only later, there are novel experiments to support special relativity in physics. Therefore, IR should not over-emphasize theories and models, as experiments are also important too. Likewise, IR should not over-emphasize in requiring experimental work in a paper, for claiming that IR is an empirical science, as some important theoretical work may have no experimental support at the time. Like special relativity, experimental support may come later after the theoretical paper is published. Similarly, heuristics are also important provided they have wide empirical support because they may later lead to some theory or model that explain or derive them. Also, some may over emphasize the importance of novelty, rejecting some scientific papers that confirm some existing theory or model, as such works have little novelty. However, these works are important as a check and balance of scientific claims. Otherwise, we may face a reproducibility crisis (Baker 2016) as in some fields of science.

After showing how IR is science, we are in a position to use the same methodology to show other subjects to be science or not science. This would involve showing the knowledge structure, the aim, the principles and the assumptions of the concerned discipline are similar to those of a known scientific discipline (e.g., physics or biology) in order to claim that the concerned scientific discipline is a science. The reasons why such a concerned scientific discipline is science will be the same as why IR is science. Therefore, we have a uniform methodology to show how subjects are science, and our ingenuity should be focused in other areas like showing the knowledge can be arranged into a knowledge structure similar to science or demonstrating how the aim of scientific study can be applied to the concerned discipline (e.g., Computer Science or Engineering Science).

Understanding IR is science enables us to draw analogy with other science subjects better. This can inspire cross-fertilization of ideas between different scientific disciplines. An existing example of cross-fertilization of ideas is between Quantum Physics and IR. For example, the Quantum PRP (Zuccon et al. 2009) is formulated for IR to take into the account of interference that is absent in the traditional PRP. Apart from principles, quantum retrieval models are also developed, like the quantum language model (e.g., Sordoni et al. 2013). Accordingly, the missing link is between Quantum PRP and Quantum retrieval model as the link is essential to mirror mature physics knowledge structures.

Finally, it may appear that any subject can be a science subject according to Luk (2017). However, Luk (2017) would consider, for example, philosophy to be not a science subject (see Table 4 for other examples). First, the aim of philosophy is not about creating scientific knowledge in the forms of theories, models and experiments. While some may regard philosophy as producing theories, philosophy does not in general create models. Traditional philosophy does not carry out any experiment. However, a new area of philosophy called experimental philosophy do carry out experiments to observe the opinions of people on philosophical topics. However, this is still far from the scientific enterprises that use the theory to construct models which are used to predict the outcomes in the experiments. The knowledge in philosophy is typically presented as arguments instead of theories, models and experiments. Second, philosophy according to Rapaport (2019) is the "personal search for truth in any field by rational means". Here, the aim is truth, which requires the accuracy of scientific knowledge to be 100% which may not be possible for some science subjects. While scientific realism may claim that a theory is true, it does not claim that all scientific theories are true. Also, scientific realism acknowledges that the scientific theories are fallible. By contrast, Luk's theory of scientific study does not require scientific knowledge to be 100% accurate. Instead, it should be as high as can be achieved by humanity and statistically better than random guess. Note that philosophy is a "personal search" whereas scientific study according to Luk (2017) is a social learning process. Also, would doing experiments be considered as a rational means in the study of the philosophical topics?

Apart from comparing the aims, we can also see whether the principles of the theory of scientific study by Luk (2017) is adhered in philosophy. First, are all established philosophical theories falsifiable as required by the empiricism principle? Second, since philosophy does not have models, the modeling accuracy principle is not applicable and the generalization principle cannot be applied as the theory cannot generalize any models and as

Table 4 Examples of Subjects regarded as science and non-science subjects, as well as subjects not decided yet. Some subjects are not decided yet because of my limited knowledge of these subjects rather than the topics are intrinsically undecided. Note that we have excluded examples of applied science subjects

Examples of science subjects	Examples of non-science subjects	Examples of subjects to be decided
Physics	Philosophy	Economics
Chemistry	Literature criticism	Political science
Psychology	Religious studies	Anthropology

there are no models available to generalize experiments. Third, at present the experiments done by experimental philosophy are not shared with others, so it is questionable whether philosophical studies follow the investigation objectivity principle. Fourth, the experiments in the experimental philosophy are about the opinions of the people about the philosophical topics rather than the experiment of the phenomenon that is described and explained by the philosophical theories. So, it is questionable whether experimental philosophy has scientific experiments that directly applies to the physical situations that the philosophical theories explain. If we discount those experiments in experimental philosophy as scientific experiments, then the objective experiment principle and the reliability principle cannot be applied to the philosophical experiments. Given these differences, we do not consider that philosophy is science. Having said that, many philosophical studies may be developed into some kind of science later because researchers may add models and experiments in their studies, and they may make models to predict the outcomes in experiments so that the boundary between philosophy and science is blurred. This explains why many fields of philosophical inquiries may turn out to be science subjects later.

6 Related Work

If computer science is a clear-cut science subject, then IR considered as a sub-discipline of computer science implies that IR is science. However, claiming computer science is science turns out to be more complicated as Rapaport (2019) exploring the philosophy of computer science has shown. For example, Denning (2005, 2007, 2013) and others (Gonzalo 2010; Cerf 2012) have written a number of papers trying to claim that computer science is science. However, many in the blogs (e.g., Raza 2014) think computer science is a branch of mathematics or engineering. Thus, computer science as a science does not seem to gain widespread acceptance especially for those not in the computer science field. As we cannot rely on computer science to imply that IR is science, we need to justify why IR is science. Likewise, we cannot rely on library and information science (LIS) to justify IR is science because even though IR is also a sub-discipline of LIS, we cannot find any paper that justifies why LIS is a science. However, we can consider that IR being a science is one piece of evidence supporting LIS and computer science as a science. As LIS and computer science are very broad subjects (in which IR is only an application rather than some fundamental process of LIS or computer science), some may question whether IR as a piece of evidence provides adequate support that LIS and computer science are sciences. Further work is needed to convince the skeptics on this issue. Similarly, one may wonder whether our examples or pieces of evidence supporting that IR is science are adequate. In this case, we use multiple examples or pieces of evidence to support our claim rather than just one piece of evidence. Also, the examples or pieces of evidence are supporting the fundamental process of IR scientific study, so we are more certain that IR is science.

Fuhr (2012) has stated that IR is an engineering science in the title of his speech for the Gerard Salton award lecture, but he cleverly avoided saying what engineering science is. Instead, he focused on what we should do to make the subject more science like. For example, he discussed that we should answer the why questions more rather than look at extensively the how questions. However, he did not explain why answering the why questions more would make the discipline more scientific. He is assuming that everybody knows that science is about knowledge and the quest is to understand. However, Luk (2018) argued that science may not be about understanding in terms of everyday-experience or based on intuition, as the subject matter may be counter-intuitive. Instead, the ability to have good predictions is a mandatory requirement of scientific knowledge. Nevertheless, he mentioned that understanding in the technical sense is still possible but not necessarily in laymen terms. Thus, it is not certain whether posing more why questions would make the discipline more scientific.

ACM has a banner to advance computing as a science and as a profession. It is, however, unknown how ACM defines science. Certainly, its members and fellows should tell us why computing is a science and Denning (2005, 2007, 2013) has been attempting to do this, although it is unknown whether people will be convinced that computing is science. In the past, there are many attempts to define science in laymen's terms by stating the definition of science. However, this usually over-generalizes science or over-specializes science. Philosophers of science have not relied on these definitions to define science because they heavily criticized such definitions. So, they are not discussed or used here.

The scientific method (SM) can show how IR is science by showing that IR studies conform to it. This can be done by showing that IR theory makes predictions as in here (Sect. 3.2), and that IR is empirical, which has experiments and hypotheses as in SM. The other activities in SM are not mentioned here because they are general activities (like analysis or question formulation). In summary, both SM and the work by Luk (2010, 2017) can demonstrate IR is science, but SM has been heavily criticized by philosophers (e.g., Cartwright 1995) and scientists (e.g., Cleland 2001). For examples, the SM was criticized to be different depending on who defined the SM. It was criticized for over-generalization, which includes other disciplines like engineering performing trial-and-error experiments. It was criticized for overspecialization, which excludes historical science by stressing the reproducibility in experiments. Also, it was considered a false idealization of how scientists investigate. Therefore, we avoided to use the SM alone to show IR is science.

7 Conclusion and Future Work

Science regarded as a class of subjects is a body of scientific knowledge, which consists of theories, models and experiments according to Luk (2010). Logically, to show that a subject is science the subject needs to have at least a similar knowledge structure as other scientific subjects that are well known to belong to science, like physics. Therefore, we show how some IR knowledge structure can map to Newtonian mechanics in order to demonstrate that IR is science. Apart from the knowledge structure, we also showed that the scientific study of IR is also constrained similarly to other scientific study because such IR study follows the aim, the principles and the assumptions specified in the theory of scientific study (Luk 2017). Therefore, there is reason to believe that IR study will be similar to other scientific investigations as they are similarly constrained and targeted. Since both the static aspect (i.e., science as a class of subjects) and the dynamic aspect (i.e., the scientific investigation) generating or revising the static aspect are similar to other science subjects, we believe that IR is a science.

This paper is a first attempt trying to show in what sense IR is science. Perhaps, this will not be the last attempt. However, this paper encourages IR researchers in the field to be more aware why IR is science, as well as what to do to make IR more similar to a mature scientific discipline. It also helps to explain to non-IR researchers and indeed laymen why IR is regarded as science. In the broader perspective, this is one-step towards showing why computer science is science, which is difficult given that computer science is a very broad subject. However, this would be a worthwhile endeavor as ACM is promoting computing as science. In addition, there is a lot of work to be done in IR to make the discipline more completely scientific. Lastly, we may further this study to quantify the similarity between IR and other scientific discipline (like physics) in terms of, for example, the number of characteristics they share but that is treated as a future work in here as we need to give a qualitative account in order to answer the "why" question (as Fuhr 2012 suggested) first.

Acknowledgements I thank Dr. Edward Dang for running the random search model. I also thank the anonymous reviewers for their constructive, insightful comments.

Compliance with Ethical Standards

Conflict of interest The corresponding author states that there is no conflict of interest.

References

Al-Maskari, A., Sanderson, M., & Clough, P. (2008). Relevance judgments between TREC and non-TREC assessors. In *Proceedings of the 31st ACM SIGIR conference* (pp. 683–684).
Azzopardi, L., & Roelleke, T. (2007). Explicitly considering relevance within the language modeling framework. In *Proceedings of the 1st international conference on theory of information retrieval* (pp. 125–134).
Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature, 533*(7604), 452–454.
Basat, R. B., Tennenholtz, M., & Kurland, O. (2015). The probability ranking principle is not optimal in adversarial retrieval settings. In *Proceedings of ICTIR'15* (pp. 51–60).
Cartwright, N. (1995). False idealization: A philosophical threat to the scientific method. *Philosophical Studies, 77*(2–3), 339–352.
Cerf, V. G. (2012). Where is the science in computer science? *Communications of the ACM, 55*(10), 5.
Chalmers, A. F. (2013). *What is this thing called science?*. Maidenhead: Open University Press.
Cleland, C. E. (2001). Historical science, experimental science and the scientific method. *Geology, 29*(11), 987–990.
Cooper, W. S. (1995). Some inconsistencies and misidentified modeling assumptions in probabilistic information retrieval. *ACM Transactions on Information Systems, 13*(1), 100–111.
Costa, A., & Roda, F. (2011). Recommender systems by means of information retrieval. In *Proceedings of WIMS'11*, Article no. 57.
Croft, W. B., Metzler, D., & Strohman, T. (2010). *Search engines: Information retrieval in practice*. Upper Saddle River, NJ: Pearson Addison-Wesley.
Damessie, T. T., Nghiem, T. P., Scholer, F., & Culpeper, J. S. (2017). Gauging the quality of relevance assessments using inter-rater agreement. In *Proceedings of the 40th ACM SIGIR conference* (pp. 1089–1092).
Dang, E. K. F., Wu, H. C., Luk, R. W. P., & Wong, K. F. (2009). Building a framework for the probability ranking principle by a family of expected weighted rank. *ACM Transactions on Information Systems, 27,* 4.
Denning, P. J. (2005). Is computer science science? *Communications of the ACM, 48*(4), 27–31.
Denning, P. J. (2007). Computing is a natural science. *Communications of the ACM, 50*(7), 13–18.
Denning, P. J. (2013). The science in computer science. *Communications of the ACM, 56*(5), 35–38.
Feyeraband, P. (2011). *The tyranny of science*. London: Polity Press.
Fuhr, N. (2008). A probability ranking principle for interactive information retrieval. *Information Retrieval, 11*(3), 251–265.
Fuhr, N. (2012). Salton award lecture information retrieval as an engineering science. *ACM SIGIR Forum, 46*(2), 19.
Fuhr, N. (2017). Some common mistakes in IR evaluation, and how they can be avoided. *ACM SIGIR Forum, 51*(3), 32–41.
Gonzalo, G. (2010). Is computer science truly scientific? *Communications of the ACM, 53*(7), 37–39.
Greiff, W. R. (1998). A theory of term weighting based on exploratory data analysis. In *Proceedings of the 21st ACM SIGIR conference* (pp. 11–19).

Huston, S., & Croft, W. B. (2014). A comparison of retrieval models using term dependencies. In *Proceedings of the 23rd ACM CIKM conference* (pp. 111–120).
Indri. (2013). *INDRI: Language modeling meets inference networks*. The Lemur Project. Retrieved June 27, 2020 from http://lemurproject.org/indri/.
Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information System, 20*(4), 422–446.
Kosso, P. (2007). Scientific understanding. *Foundations of Science, 12*(2), 119–130.
Lafferty, J., & Zhai, C. X. (2001). Probabilistic relevance models based on document and query generation. In B. Croft & J. Lafferty (Eds.), *Language modeling for information retrieval* (pp. 1–10). Dordrecht: Springer.
Lavrenko, V. (2009). *A Generative Theory of Relevance*. Berlin: Springer.
Lin, J. (2018). The neural hype and comparison against weak baselines. *ACM SIGIR Forum, 52*(2), 40–51.
Luk, R. W. P. (2008). On event space and rank equivalence between probabilistic retrieval models. *Information Retrieval, 11,* 539–561.
Luk, R. W. P. (2010). Understanding scientific study via process modeling. *Foundations of Science, 15*(1), 49–78.
Luk, R. W. P. (2017). A theory of scientific study. *Foundations of Science, 22*(1), 11–38.
Luk, R. W. P. (2018). To explain or to predict: Which one is mandatory? *Foundations of Science, 23*(2), 411–414.
Maron, M. E., & Kuhns, J. L. (1960). On relevance, probabilistic indexing and information retrieval. *Journal of the ACM, 7*(3), 216–244.
Paik, J. H. (2013). A novel TF-IDF weighting scheme for effective ranking. In *Proceedings of the 36th ACM SIGIR conference* (pp. 343–352).
Popper, K. (1959). *The logic of scientific discovery*. London: Hutchinson.
Rapaport, W. J. (2019). *Philosophy of computer science*. Retrieved March 25, 2019 from http://cse.buffalo.edu/~rapaport/Papers/phics.pdf.
Raza, K. (2014). *Is the discipline "computer science" a "natural science"?* Retrieved June 27, 2020 from https://www.researchgate.net/post/Is_the_discipline_Computer_Science_a_Natural_Science2.
Reiss, J., & Sprenger, J. (2017). Scientific objectivity. In E. N. Zalta (Eds.), *The Stanford encyclopedia of philosophy* (Winter 2017 Edition). Retrieved June 27, 2020 from https://plato.stanford.edu/archives/win2017/entries/scientific-objectivity.
Robertson, S. E. (1977). The probability ranking principle in IR. *Journal of Documentation, 33,* 294–304.
Robertson, S. E. (2006). On GMAP: And other transformations. In *Proceedings of the 15th ACM CIKM conference* (pp. 78–83).
Saracevic, T. (1975). Relevance: A review of and a framework for the thinking on the notion in information science. *Journal of the Association for Information Science and Technology, 26*(6), 321–343.
Singhal, A., Buckley, C., & Mitra, M. (1996). Pivoted document length normalization. In *Proceedings of the 19th ACM SIGIR conference* (pp. 21–29).
Sordoni, A., Nie, J.-Y., & Bengio, Y. (2013). Modeling term dependencies with quantum language models for IR. In *Proceedings of the 36th ACM SIGIR conference* (pp. 653–662).
Spärck-Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation, 28*(1), 11–21.
Terrier. (2019). *Terrier v5.1*. University of Glasgow. Retrieved July 3, 2019 from http://terrier.org.
Van Fraassen, B. (1980). *The scientific image*. Oxford: Clarendon Press.
Van Rijsbergen, C. J. (1979). *Information retrieval*. London: Butterworths.
Van Rijsbergen, C. J. K. (2006). Quantum haystacks. In *Proceedings of the 29th ACM SIGIR conference* (pp. 1–2).
Wong, K. F., Song, D., Bruza, P., & Chen, C.-H. (2001). Application of aboutness to functional benchmarking in information retrieval. *ACM Transactions on Information Systems, 19*(4), 337370.
Wu, H. C., Luk, R. W. P., Wong, K. F., & Kwok, K. L. (2008). Interpreting TF-IDF weights as making relevance decisions. *ACM Transactions on Information Systems, 26,* 3.
Yang, P, & Feng, H. (2016). A reproducibility study of information retrieval models. In *Proceedings of ICTIR'16* (pp. 77–86).
Zamani, H., Croft, W. B., & Culpepper, J. S. (2018). Neural query performance prediction using weak supervision from multiple signals. In *Proceedings of the 41st ACM SIGIR conference* (pp. 105–114).
Zhai, C. X. (2011). Axiomatic analysis and optimization of information retrieval models. In *Proceedings of ICTIR 2011 conference* (p. 1).
Zhai, C. X., & Lafferty, J. (2004). A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems, 22*(2), 179–214.

Zobel, J. (2017). What we talk about when we talk about information retrieval. *ACM SIGIR Forum, 51*(3), 18–26.
Zuccon, G., Azzopardi, L. A., & van Rijsbergen, C. J. K. (2009). The quantum probability ranking principle for information retrieval. In *Proceedings of the ICTIR '09* (pp. 232–240).
Zuo, J., Wang, M., Wan, J., Wu, G., & Wu, S. (2012). Modified information retrieval model based on Markov network. In *Proceedings of international conference on network computing and information security* (pp. 307–314).

Robert W. P. Luk received a B.Sc. degree, a Dip. Eng. and a Ph.D. degree from the School of Electronics and Computer Science, University of Southampton, and M.Sc. degree from the Department of Psychology, University of Warwick. He is a visiting research scholar at the Center for Intelligent Information Retrieval, University of Massachusetts, Amherst, USA. He is a PC member of various conferences (e.g., ACM SIGIR) and is a reviewer of various journals (e.g., ACM Trans. Asian Language Information Processing and IEEE Trans. Systems. Man and Cybernetic). He is an associate professor of the Department of Computing, The Hong Kong Polytechnic University.