

Gravitation-Based Model for Information Retrieval

Shuming Shi, Ji-Rong Wen, Qing Yu¹, Ruihua Song, Wei-Ying Ma

Microsoft Research Asia, 49 Zhichun Road, Beijing, 100080, P.R. China

{shumings, jrwen, rsong, wyma}@microsoft.com

¹Department of Computer Science, Beijing Institute of Technology, Beijing, P.R. China

¹f-qyu@mrschina.research.microsoft.com

ABSTRACT

This paper proposes GBM (*gravitation-based model*), a physical model for information retrieval inspired by Newton's theory of gravitation. A mapping is built in this model from concepts of information retrieval (documents, queries, relevance, etc) to those of physics (mass, distance, radius, attractive force, etc). This model actually provides a new perspective on IR problems. A family of effective term weighting functions can be derived from it, including the well-known BM25 formula. This model has some advantages over most existing ones: First, because it is directly based on basic physical laws, the derived formulas and algorithms can have their explicit physical interpretation. Second, the ranking formulas derived from this model satisfy more intuitive heuristics than most of existing ones, thus have the potential to behave empirically better and to be used safely on various settings. Finally, a new approach for structured document retrieval derived from this model is more reasonable and behaves better than existing ones.

Categories and Subject Descriptors

H.3.3 [Information Search and Retrieval]: Retrieval models

General Terms

Algorithms, Experimentation, Theory

Keywords

Information retrieval models, Gravitation-based model, theory of gravitation, mass estimation, structured document retrieval

1. INTRODUCTION

Information retrieval (IR) models, which define the representation of documents, queries, and the relevance relationship between them, are in a core position in information retrieval (IR). In the past several decades, many categories of IR models (and their variants) have been proposed and studied [2], including Boolean

models, vector space models [3][4], probabilistic and logic models [10][14][6][1], and language models [12][13][7][24], etc. The key behind all the models is the primary *perspective* on information retrieval. The Boolean model views IR problems from the perspective of set theory and Boolean algebra, while the perspective used in the vector space model is vector and linear algebra. Most of other categories of models take the probabilistic perspective, which is the most dominating perspective on information retrieval today.

It may be extremely hard to answer questions like "what is the essence of information retrieval", and "what is the right perspective of it". However, it is clear that, till now, we know more about information retrieval each time when a new perspective is adopted. It would also be helpful to view information retrieval from more new perspectives.

Although many of the models (and the formulas and algorithms derived from them) have been successfully applied to various tasks, there are still some problems faced by them: First, the retrieval formulas (formal or ad-hoc) conducted by most IR models fail to satisfy even some basic intuitive heuristic constraints [5]; Second, the retrieval formulas derived or motivated from many IR models commonly lack intuitive interpretations, especially *physical* interpretations. At the same time, we are living in a physical world which is dominated by fundamental physical laws. Can we get help from "the God" in acquiring deeper understanding of information retrieval?

In this paper, we try to view information retrieval from the perspective of physics, a quite different perspective from existing ones. We propose a new framework which models documents, queries, and their relationships using basic concepts in physics. In particular, documents and queries are modeled as objects with specific structures; and the relationship between a query and a document is modeled as the attractive force between them. A basic rule used here is Sir Isaac Newton's *theory of gravitation* (see Section 2.1 for a brief introduction of it), a fundamental law of the universe. The primary goal of the model is to help learning more about information retrieval from a new perspective.

It is encouraging that we can really benefit from the nature. With the new perspective and model, we get the following preliminary achievements,

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SIGIR'05, August 15-19, 2005, Salvador, Brazil.

1. We have derived a family of effective ranking formulas which satisfy all the heuristic constraints¹ proposed in [5]. Experimental results show that these formulas are among the most effective ranking functions proposed till now.
2. The BM25 term weighting function [9][11] can be easily derived from our basic model, so we give an intuitive *physical* interpretation of this powerful and robust function.
3. A more reasonable approach for structured document retrieval can be obtained from the model. This approach is not only highly effective but also robust to be used in various conditions.

In this paper, we will examine the gravitation-based model theoretically and empirically. We first give some background knowledge and related work in section 2. In section 3, the GBM model will be introduced and analyzed theoretically. The performance of the model is tested by some experiments in section 4. We conclude the paper in Section 5.

2. BACKGROUND AND RELATED WORK

In this section, we will first give a brief introduction of Newton's theory of gravitation, upon which our model is built. And then some related work is discussed.

2.1 Newton's Theory of Gravitation

Gravitation is one of the four fundamental forces of nature. It governs the motion of stars and plays a crucial role in most processes on the earth. Newton proposed in 1687 his famous law of gravitation which demonstrated that any two objects in the universe have attractive force between them. For two particles with masses m_1 , m_2 respectively and with distance d between them, the gravitational force between them can be expressed as follows,

$$F = \frac{Gm_1m_2}{d^2} \quad (2.1)$$

where G is a constant called the universal gravitation constant. And the direction of the force is along a line between the two particles. It can also be proved by calculus that the gravitational force between two spheres can be viewed as all their masses are concentrated in their centers.

2.2 Information Retrieval Perspectives and Models

All IR models have their primary perspectives on information retrieval problems. Some ranking formulas for retrieval are commonly derived or motivated from the models. The *term weighting function*, which defines the score of a document given one query term, is the most important part of a ranking formula.

The followings are a brief overview of some categories of IR models and the most effective term weighting functions for them.

2.2.1 Vector space model

In the vector space model [3][4], each document is represented as a vector of terms, so does a query. And the relevance is measured by the similarity (e.g. cosine of angle) between the query vector and the document vector. Many term weighting functions have been proposed upon this model (and its variants), among which the pivoted normalization weighting formula [4] seems to be an outstanding one,

$$w(D, Q, t) = \frac{1 + \ln(1 + \ln(c(t, D)))}{(1-s) + s \frac{|D|}{avdl}} \cdot \ln \frac{N+1}{df(t)} \quad (2.2)$$

where s (between 0.0 to 1.0) is a parameter. Please see Table 1 (section 3.1) for the notations used in the above formula.

2.2.2 Probabilistic model

The probabilistic model [10][14][6][1] formulates the IR problem in a probabilistic framework, which gives the relevance of a document and a user query by estimating the probability that the document is exactly what the user needs. Variants of probabilistic models include Bayesian networks, inference network models, belief network models, etc. Please refer to [8] for an overview of them.

Okapi's BM25 formula [9][11] is shown as one of the most effective and robust ranking formulas in this category (and even in all formulas till now). The term weighting function of its commonly used simplification [9][11][16] is,

$$w(D, Q, t) = \frac{(k_1 + 1) \cdot c(t, D)}{k_1 \cdot ((1-b) + b \cdot \frac{|D|}{avdl}) + c(t, D)} \cdot w(t) \quad (2.3)$$

Where k_1 and b are parameters. The origin representation of $w(t)$ has the potential "negative IDF" issue, as has been discussed in [5]. Like in [5], we will use $\ln((N+1)/df(t))$ as the expression of $w(t)$ in the following part of the paper.

The above formula is first discovered by Robertson et al [9][10][11], inspired by the *shape* of a complex formula derived from a probabilistic model under the 2-Poisson assumption. Amati and Rijsbergen propose in [1] a probabilistic framework for generating nonparametric term weighting functions. They claim that the BM25 function with some special parameters ($k_1=1.2$, $b=0.75$; or $k_1=2$, $b=0.75$) can be *approximated* numerically by one of their generated functions $I(n)L2$ (with $k_1=1$ and 2 respectively). By following these brilliant works, we try to derive BM25 from our model and give it a physical explanation.

2.2.3 Language model

The language model [12][13][7][24] also adopts a probabilistic framework. However, different from traditional probabilistic models, it interprets the relevance between a document and a query as the probability of generating the query from the document's model. Smoothing, which adjusts term probabilities to

¹ There is a small issue for the TDC constraint, which will be discussed in Section 3.2.4.3.

overcome data sparseness, is critical to the performance of language models. Among various smoothing methods, the Dirichlet prior smoothing seems to be discussed frequently,

$$w(D, Q, t) = \ln \left(I \cdot \frac{c(t, D)}{|D|} + (1 - I) \cdot P_{MLE}(t | C) \right) \quad (2.4)$$

where $I = |D| / (|D| + m)$, and $P_{MLE}(t | C)$ is the maximum likelihood estimate of the probability of term t in collection C . And m is a parameter whose value is commonly set to be multiples of the average document length.

2.3 Structured Document Retrieval

As we will discuss in Section 3.3, our model can support structured document retrieval naturally and effectively. So another kind of work related to ours is structured document retrieval. A document is said to be structured when it contains multiple fields. Document's field structure is commonly used to improve retrieval performance in practice.

The most commonly used approach for structured document retrieval may be score/rank (linear) combination [15][18][19][20], which treats each field as a separate document and computes scores/ranks for them. In computing scores for each field, any ranking function for unstructured document retrieval can be adopted.

Robertson et al [16] introduced, as an extension of the BM25 formula, a simple and efficient method which combines term frequencies instead of field scores. Ogilvie et al [17] proposed, within the language model framework, another approach which develop a separate language model for each field and then fix them linearly. These approaches essentially combine term frequencies instead of scores.

We will discuss some issues with the above (two kinds of) methods and their relations with ours in Section 3.3.

3. GRAVITATION-BASED MODEL

The gravitation-based model (GBM) tries to understand the IR problem within the physical framework. To achieve this, we first build a mapping from the concepts in information retrieval to those in physics (Section 3.1). And then a basic model (with two versions: discrete and continuous) is introduced and analyzed (Section 3.2). Then, in Section 3.3, we will show how GBM supports structured document retrieval naturally and effectively. Finally, two possible future extensions for GBM are briefly discussed.

Table 1. Commonly used IR notations

Expression	Description
$ D $	The length (number of terms) of document D
$avdl$	Average document length in a collection
$df(t)$	Document frequency (DF) of term t
$c(t, D)$	Times of occurrences of term t in document D
N	The number of documents in current collection

Table 2. Notations only for the GBM model

Expression	Description
$m(t, D)$	Mass of term t in document D
$m(t, Q)$	Mass of term t in query Q
$\bar{m}(t)$	Average mass of term t in the collection
$m(D)$	Mass of document D
$di(t, D)$	The diameter of term t in document D
$di(D)$	The diameter of document D
$r(t, D)$	The radius of term t in document D
$H(D)$	The set of hidden terms in document D

3.1 Notations and Basic Concepts

As in some other IR papers, we use D , Q , C to denote a document, a query, and the whole collection respectively. Some other commonly used notations are listed in Table 1. In the GBM model, a term (word) in a document is viewed as a physical object composed of some amount of particles.

3.1.1 Particle

A particle has two attributes with it: type, and mass. The type of a particle is determined by the term it composes. For different occurrences of the same term, their corresponding particles have the same type. Two particles of the same type have some kind of gravitational force between them and so they are mutually attracted. But there is no any force between particles of different types. For two particles P_1, P_2 with mass m_1 and m_2 respectively, the gravitational force between them is (by formula 2.1),

$$F(P_1, P_2) = \begin{cases} \frac{Gm_1m_2}{d^2} & \text{if } type(P_1) = type(P_2) \\ 0 & \text{otherwise} \end{cases} \quad (3.1)$$

where d is the distance between the two particles, and G is a constant.

3.1.2 Term object

A term is an object composed of particles with a specified structure (or shape). Two shapes (see Figure-1) will be discussed in this paper: the sphere, and the *ideal* cylinder. An *ideal* cylinder is a cylinder whose radius is small enough and so that it can be viewed as a line segment.

There are three attributes related to a term: type, mass, and diameter. For a sphere-shape term, its diameter is defined naturally as the sphere's diameter. While for a term with an ideal cylinder shape, its diameter is defined as its height (see Figure 1(b)). The mass and diameter of a term t in document D are denoted as $m(t, D)$ and $di(t, D)$ respectively. Also define for each term a radius (denoted by $r(t, D)$) as the half value of the diameter (see Table-2 for a list of notations).

The two kinds of term shapes will be used respectively by the discrete and continuous versions of GBM (Section 3.2).

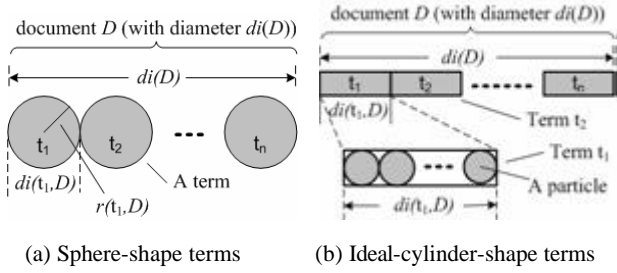


Figure 1. Term and document objects in the GBM model

3.1.3 Document object

A document is modeled as a higher level object (than a term) composed of a list of terms objects. There are two categories of term objects in each document: One category, called *explicit* (or seen) terms, includes all the terms occurred in the content of the document. The other category is called *implicit* (or hidden) terms which are not occurred in the document. Implicit terms are used to represent the *hidden* meaning of the document. All the hidden terms in document D are denoted as $H(D)$. We use $|D|$ and $|H(D)|$ to represent the number of explicit and implicit terms of D respectively. As a result, the total number of terms of the whole document can be expressed as $|D|+|H(D)|$, or $|D \cup H(D)|$.

It is reasonable and natural to assume that the mass of a document is the total masses of all its seen and hidden terms,

$$m(D) = \sum_{t \in D \cup H(D)} m(t, D) \quad (3.2)$$

The fundamental logic behind is that the idea of a document is expressed by and only by all its seen and hidden terms.

As a document is modeled as a *list* of terms, we can also define a *diameter* for it. The diameter of a document is (naturally) defined as the sum of the diameters of all its terms. That is,

$$di(D) = \sum_{t \in D \cup H(D)} di(t, D) \quad (3.3)$$

3.1.4 Query object

Similar to the modeling of a document, a query is modeled as an object composed of its terms (no hidden terms are assumed for simplicity). Assume that each query has unit mass².

3.1.5 Relevance as gravitational force

The relevance of a document given a query is modeled as the force between the objects corresponding to them. According to physics principles, the force between two objects is determined by their shapes, their masses, and the distance between them.

3.2 The Basic Model

Here we describe two basic versions of gravitation-based model in which only TF, IDF, and document length is considered. In the discrete version, terms are modeled as spheres; while in the continuous version, each term is represented as an ideal cylinder.

3.2.1 A discrete version

Although a document by default is a list of terms sorted by their natural order in the document, its structure is assumed to be changed under the attraction of a query in this model, as is illustrated in Figure 2. In the figure, spheres labeled with “ t_1 ”, “ t_2 ”, etc are the occurrences of *query terms* in the document, while those labeled with “ x ” are other terms (i.e. terms that are not in query Q).

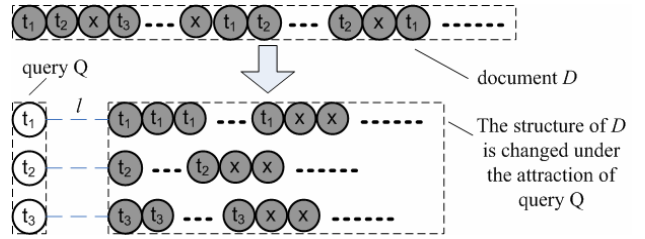


Figure 2. Discrete version of the basic GBM model

Under the gravitational attraction of query Q , the terms which feel larger gravitational forces will get nearer to the query. Given a query, each document has an optimized term placement where the aggregated force between the query and the document is maximal. In this model, the relevance of a document given a query is defined by the attractive force between them when the document is in its optimal-term-placement state.

Now we have transformed the relevance computation in IR into a term placement optimization problem. It is clear from Figure 2 that the maximal (optimized) gravitational force between a query term t and document D (using formula 3.1) is,

$$F_{gbm}(t, D) = \sum_{i=0}^{c(t, D)-1} \frac{G \cdot m(t, D)}{(l + (i + \frac{1}{2}) \cdot di(t, D))^2} \quad (3.4)$$

Where l is the distance between the query and the document (see Figure 2). According to the principle of the composition of forces, the force between D and Q can be expressed as follows,

$$F_{gbm}(Q, D) = \sum_{t \in Q} F_{gbm}(t, D) \quad (3.5)$$

Therefore we get the expression of relevance between D and Q .

3.2.2 A continuous version

Here we introduce the continuous version of our model, by representing terms as ideal cylinders of particles (see Section 3.1.2 and Figure 1). The relevance of a document given a query is still defined by the attractive force with optimal placement (the same as in the discrete one). It is clear from Figure 3 that the attractive force between a query term t and document D can be expressed as follows (by using formula 3.1),

$$F_{gbm}(t, D) = \int_l^{l+c(t, D) \cdot d(t, D)} \frac{G \cdot m(t, D)}{di(t, D) \cdot x^2} dx = \frac{G \cdot m(t, D)}{l \cdot di(t, D)} \cdot \frac{c(t, D)}{\frac{l}{di(t, D)} + c(t, D)} \quad (3.6)$$

² See our technique report [21] for a detail discussion about it.

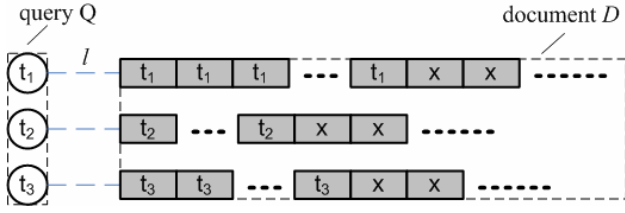


Figure 3. Continuous version of the basic GBM model. Document D is in an optimized term-placement state.

The above formulas (3.4 and 3.6) can not be used directly for retrieval because the values of $m(t,D)$ and $di(t,D)$ are unknown. We will (in Section 3.2.4) deduce the ultimate formula for retrieval after making estimations (in Section 3.2.3) for the mass and diameter of a term in a document.

3.2.3 Mass and diameter estimation

To simplify the model and make the deduction process concise, we give some *simplicity* assumptions on which the estimation is based. Please note that the model itself does not rely on these assumptions.

As documents vary in their masses and lengths, the same term may have different mass values in different documents. However, we can define a document-independent mass for each (type of) term, $\bar{m}(t)$, which denotes the average mass of term t in the whole collection. We can also consider $\bar{m}(t)$ as the global importance of term t in the considered collection. We then simply make the following simplicity assumptions,

Assumption 1 For any two terms, their mass ratio in any document is equal to the ratio of their average masses in the whole collection. This can be formally expressed as

$$\frac{m(t_1, D)}{\bar{m}(t_1)} = \frac{m(t_2, D)}{\bar{m}(t_2)} \xrightarrow{\text{define as}} h(D) \quad (3.7)$$

Assumption 2: The average mass of a term depends on and only on the inverse document frequency. Using the IDF expression in Formula 2.2, we have

$$\bar{m}(t) = \ln((N+1)/df(t)) \quad (3.8)$$

Assumption 3: The average global importance of all terms in a document is constant (i.e. independent of the document itself). This assumption implies that most documents are less likely to contain only weighty or light terms, but have more chance to contain terms with various masses. Formally,

$$\frac{\sum_{t \in D \cup H(D)} \bar{m}(t)}{|D \cup H(D)|} = c_1 \quad (3.9)$$

Now we try to estimate the mass of a term in a document. By combining formula 3.2, 3.7, and 3.9, we get,

$$m(t, D) = \frac{m(D)\bar{m}(t)}{c_1 \cdot (|D \cup H(D)|)} \quad (3.10)$$

The following assumptions are for estimating the diameter of a term in a document.

Assumption 4: The number of hidden terms in a document relies only on the statistic information of the whole collection, while is independent of the document itself. That is,

$$|H(D)| = c_2 \cdot avdl \quad (3.11)$$

As $|D \cup H(D)| = |D| + |H(D)|$, we have the following formula,

$$|D \cup H(D)| = (1 + c_2) \cdot avdl \cdot \left((1 - \mathbf{b}) + \mathbf{b} \frac{|D|}{avdl} \right) \quad (3.12)$$

where $\mathbf{b} = 1/(1+c_2)$.

Also assuming all terms in the same document have equal diameters for simplicity and according to formula 3.3, we have,

$$di(t, D) = \frac{di(D)}{|D \cup H(D)|} = \frac{di(D)}{|D| + c_2 \cdot avdl} \quad (3.13)$$

3.2.4 Model analysis

Now we analyze the proposed model using the mass and diameter estimation results in the previous sub-section. For the continuous model, by applying formula 3.10 and 3.13 to formula 3.6, we have,

$$F_{gbm}(t, D) = \mathbf{r}(D) \cdot \bar{m}(t) \cdot \frac{c(t, D)}{\frac{l}{\mathbf{e}(D)} \cdot \mathbf{j}(D) + c(t, D)} \quad (3.14)$$

where $\mathbf{r}(D) = \frac{G \cdot m(D)}{c_1 \cdot l \cdot di(D)}$, $\mathbf{e}(D) = \frac{di(D)}{(1+c_2) \cdot avdl}$, and

$\mathbf{j}(D) = (1 - \mathbf{b}) + \mathbf{b} \cdot \frac{|D|}{avdl}$. And the physical meaning of $\mathbf{e}(D)$ is the

diameter per term when document D have average document length.

Similarly, formula 3.4 for the discrete model can be transformed as (by applying formula 3.10 and 3.13),

$$F_{gbm}(t, D) = \mathbf{r}(D) \cdot \bar{m}(t) \cdot \sum_{i=0}^{c(t, D)-1} \frac{l}{(l + (i + \frac{1}{2})) \cdot \frac{\mathbf{e}(D)}{\mathbf{j}(D)}} \quad (3.15)$$

Formula 3.14 and 3.15 are the final term ranking functions derived from our GBM model.

3.2.4.1 Relationship with the BM25 formula

In previous section, if $m(D)$ and $di(D)$ are constant, then expression $\mathbf{r}(D)$ and $\mathbf{e}(D)$ are constant accordingly. In this condition, it is easy to prove that formula 3.14 is equivalent to formula 2.3, the (simplified) BM25 term weighting function. Note that $l/\mathbf{e}(D)$ and \mathbf{b} in formula 3.14 are corresponding to parameter k_1 and b in formula 2.3 respectively. And $\mathbf{r}(D)$ can be ignored (because it is a constant).

Formula 3.15 also has a close relationship with BM25. Please refer to our technique report [21] for details.

3.2.4.2 A family of effective ranking formulas

So far the gravitation field function is inverse-square ($1/x^2$). However, other gravitation fields can be considered. By similar process with formula 3.14, we can get a family of formulas (see Table 3) for the continuous basic GBM model, corresponding to different force field functions respectively. Among them, the

Okapi formula (corresponding to the case of $pow=2$) is the simplest and most easy-to-compute³ one!

Table 3. Selected term weighting functions derived from GBM

Gravitational-Field-Function	Term weighting function derived ($F_{gbm}(t, D)$)
$1/x^{pow}$ ($pow \neq 1$)	$\frac{r(D)}{1-pow} \cdot \bar{m}(t) \cdot \left(1 + \frac{e(D) \cdot c(t, D)}{l \cdot j(D)}\right)^{1-pow} - 1$
$1/x$	$r(D) \cdot \bar{m}(t) \cdot \ln\left(1 + \frac{e(D)}{l \cdot j(D)} \cdot c(t, D)\right)$
e^{-x}	$r(D) \cdot \bar{m}(t) \cdot \left(1 - \exp\left(-\frac{e(D)}{l \cdot j(D)} \cdot c(t, D)\right)\right)$

3.2.4.3 Checking with heuristic constraints

In [5], Fang et al proposed some heuristic constraints related to TF, IDF, and document length that all reasonable ranking formulas should follow, namely TFC1, TFC2, TDC, LNC1, LNC2, and TF-LNC. As there is a small issue for constraint TDC⁴, so we replace it with M-TDC here.

It can be checked that each formula in Table-3 satisfies all the above heuristic constraints. The details of analysis are omitted here due to space limitations.

Experimental results (see Section 4) indicate that the formulas derived from the basic model have high performance.

3.3 Exploiting Document Structure s

In the basic GBM model, a document is treated as a bag of words, and the relevance of a documents respected to a query is examined by only considering TF, IDF, and document length. However, the power of our model is more than this. In this subsection, we show that GBM can support structured document retrieval easily, formally and in a natural way.

³ Only the four fundamental operations of arithmetic (addition, subtraction, multiplication, and division) are needed to compute it.

⁴ The TDC constraint in [5] says “Let q be a query and $w_1, w_2 \in q$ be two query terms. Assume $|d_1|=|d_2|$, $c(w_1, d_1)+c(w_2, d_1)=c(w_1, d_2)+c(w_2, d_2)$. If $idf(w_1) \geq idf(w_2)$ and $c(w_1, d_1) \geq c(w_1, d_2)$, then $f(d_1, q) \geq f(d_2, q)$ ”. Consider the following case: let $c(w_1, d_1)=10$, $c(w_2, d_1)=0$, $c(w_1, d_2)=5$, $c(w_2, d_2)=5$, and assume $idf(w_1)$ is only slightly larger than $idf(w_2)$. By applying TDC, we will have $f(d_1, q) \geq f(d_2, q)$. However, it is more reasonable and intuitive (supported by TFC1 and TFC2) that $f(d_1, q) < f(d_2, q)$. So there should be a small bug in TDC. This constraint would become reasonable if condition “ $c(w_1, d_1)+c(w_2, d_1)=c(w_1, d_2)+c(w_2, d_2)$ ” was replaced by “ $c(w_1, d_1)=c(w_2, d_2)$ and $c(w_2, d_1)=c(w_1, d_2)$ ”. With this modification, it is simply referred as M-TDC (modified TDC) here.

3.3.1 An approach derived from our model

In modeling structured documents, the fundamental principles are the same as those in the model for unstructured documents. They are 1) Terms will compete for places and the ones have larger gravitational forces will get nearer to the query terms, and 2) The relevance of a document given a query is defined by the maximal gravitational force between them among all possible term placements.

The only difference is that each field F has its mass $m(F)$, and different kinds of fields may have different masses. The mass of a term in a field F can now be computed by a formula similar with Formula 3.10, but with document D in it replaced by field F .

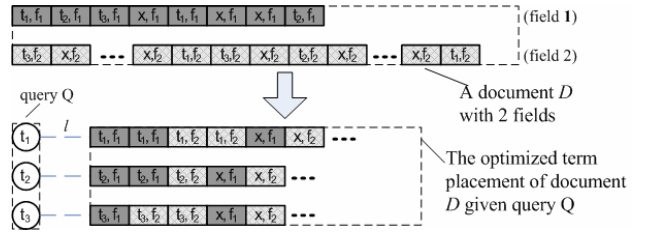


Figure 4. Support for structured document retrieval in GBM

In Figure 4 we explain how our model deals with document structures. In the figure, “ t_1, f_1 ” etc represents term t_1 in field f_1 of the document, and two kinds of filling styles for rectangles represent two fields respectively. And because masses of terms in field 1 are larger than those of field 2, field 1’s terms are nearer to the query. The optimized term placement for document D is shown in the figure. Please note that terms in different fields may have different diameters as well.

Note that this is NOT the only structured document retrieval approach that can be derived from our model. But it is in the most simplest and easy-to-understand ones.

3.3.2 Analysis and comparison

As introduced in Section 2.3, most methods for structured document retrieval adopt some kinds of combination over the fields: score/rank combination, or term frequency combination.

Robertson et al [16] argued that score combination “can be quite dangerous” and they prefer TF combination. We agree with their discussions about score combination. For a multi-term query, with score combination, a document matching a single query term over many fields could get *unreasonably* higher score than another document which matches all the query terms in a few fields. Score combination is not reasonable at this aspect.

However, the results of TF combination may also be *unreasonable* in some conditions. Please see the following example (assume the TF combination method in [16] is used):

Assume four documents d_1, d_2, d_3, d_4 with equal length, and each has two fields F_1, F_2 whose weights (assigned by a TF

combination method) are assumed to be 5 and 1 respectively. Also assume that the term frequencies of a term t in the two fields of the four documents are as follows,

$$\begin{aligned} \text{tf}(t, d_{1,f_1}) &= 1; \text{tf}(t, d_{2,f_1}) = 0; \text{tf}(t, d_{3,f_1}) = 1; \text{tf}(t, d_{4,f_1}) = 0 \\ \text{tf}(t, d_{1,f_2}) &= 0; \text{tf}(t, d_{2,f_2}) = 6; \text{tf}(t, d_{3,f_2}) = 8; \text{tf}(t, d_{4,f_2}) = 14 \end{aligned}$$

Now consider the relevance of the documents given a query containing only one term t . By TF combination, the score of d_1 is small than that of d_2 (since $1*5+0 < 0*5+6$). This may be reasonable. But if applying the same TF combination method to d_3 and d_4 , we find that $\text{score}(d_3) < \text{score}(d_4)$ (since $1*5+8 < 0*5+14$). This is not coherent to common sense, because it is clearly that the score of d_3 (which contains term t in its high-weight field, and enough times of the same term in its low-weight field) should be larger than that of d_4 (which does not contain t in its high-weighted field at all).

From the above discussion, we see that both score combination and the linear combination of TFs fail to pass the test of simple heuristic constraints. Although they may have been successfully applied to many datasets and tasks, it is really dangerous to adopt them in all conditions. We DO need some heuristic constraints for the retrieval of structured documents, just like those proposed by [5] for unstructured documents.

In contrast to the two existing methods, the approach derived from our model is immune from the above two issues. We are not saying here that our method will satisfy all the potential heuristic constraints in the future. But, as our model is originated from basic physical concepts and laws, it can be hoped to have more chances to obey human intuitions. The retrieval performance of our approach will be tested by experiments in section 4.

3.4 Possible Extensions

In this subsection, we illustrate briefly two potential extensions to the GBM model: term proximity, and the combination between relevance scores and static document ranks.

Till now, term proximity information is omitted in our model. However, it is intuitive that adjacent query terms in a document should have larger gravitational forces (with the query) than distant ones. So our model is promising to provide a clear explanation of term proximity.

In previous analysis and discusses, document mass $m(D)$ (see formula 3.14 and 3.15) is simply set to be constant. However, the mass of a document is a measure of its quality, which depends on how informative and important the document is. So document mass may have some strong relationships with static document ranks (e.g. PageRank [22] in Web search). We plan to exploit this in the future.

4. EXPERIMENTS

In this section, we verify the analysis of previous section and test the performance of our model by some experiments performed on standard datasets and queries.

4.1 Experimental Methods

We performed experiments on two corpora and seven query sets (see Table 3 and 4) used from TREC [23] 2000 to 2004.

All the experiments are performed on an information retrieval platform built by our organization. In parsing documents, image alt information is extracted as part of body-text, while meta keywords and descriptions are discarded. A naïve word breaker (which treats characters other than letters and digitals as punctuations) is adopted to separate document text as terms. All terms are indexed, i.e. no stop-words are removed. In the processing of document and query terms, the Porter stemmer is used for stemming.

Table 4. Corpora characteristic

Corpus	#Documents	Size	AVDL	Document types
WT10G	1,692,096	9.8GB	537.4	Html
.GOV	1,274,753	17.8GB	998.7	Html, pdf, ps, doc

Table 5. Query-sets used in the experiments

Query-Set	2000.w eb	2001.a dhoc	2002. np	2002. td	2003. np	2003. td	2004. mix
For corpus	WT10G	WT10G	.GOV	.GOV	.GOV	.GOV	.GOV
#Topics	50	50	150	50	300	50	225

To evaluate the query results, we use mean average precision (MAP), since it applicable to all the tasks to be performed here.

In tuning parameter to optimize an evaluation measure, we use a grid search method similar with that described by Robertson et al in [16], “evaluate the performance of a system on a successively smaller grid over the set of parameters being optimized, until an adequate minimum-step value is reached”.

4.2 Term Weighting Experiments

In this subsection, we compare the formulas derived from our basic model with other most commonly used formulas. Although quite a lot of term weighting functions have been proposed, those listed in Section 2 (see formula 2.2, 2.3, and 2.4) are recognized as most effective ones in their categories respectively. We would not like to compare our formulas with the earliest VSM or probabilistic formula or one of the language model formulas without smoothing to get a “seems-like” significant performance improvement. The formulas participated in comparison are VSM-Piv (the pivoted normalization VSM formula, see Formula 2.2), LM-Dir (the language model formula with Dirichlet prior smoothing, see Formula 2.4), GBM-Std (formula 3.14 derived from our model, which is equivalent to the BM25 formula), GBM-Std-Disc (formula 315 derived from our discrete model), GBM-Inv (the GBM formula derived by gravitational field function $1/x$, see Section 3.2.4.2), and GBM-Exp (the GBM formula derived by e^{-x}). Some of the comparison results are shown in Table 6. Only body text is used to generate the results, and we use mean average precision as the performance measure.

Table 6. Optimal performance comparison among (the considered) formulas over various corpora and tasks

	VSM-Piv	LM-Dir	GBM-Std (Okapi)	GBM-Std-Disc	GBM-Inv	GBM-Exp
2000.web	0.190	0.190	0.190	0.187	0.195	0.186
2001.adhoc	0.181	0.182	0.179	0.177	0.184	0.174
2002.np	0.534	0.55	0.554	0.54	0.545	0.542
2002.td	0.174	0.195	0.195	0.192	0.182	0.186
2003.np	0.382	0.384	0.402	0.392	0.393	0.395
2003.td	0.116	0.109	0.116	0.126	0.122	0.115
2004.mix	0.247	0.239	0.257	0.261	0.252	0.248

From Table 6, we can see that the performances of the ranking formulas listed are comparable, and GBM-Std (i.e. Okapi) performs slightly better than VSM-Piv and LM-Dir. These results give more confidence that the formulas derived from our basic model are *comparable to or better than* state-of-the-art high performance ones. In addition, as our formulas satisfy all the heuristic constraints proposed (while most of others do not), they are hoped to be robust on various kinds of datasets and tasks.

Please note that the average precision for some tasks is lower than the top results in TREC submissions. This is because only body text is used here.

4.3 Field Structure Experiments

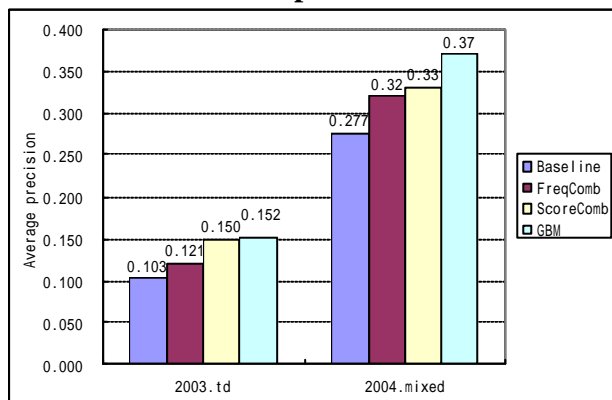


Figure 5. Performance comparison of different approaches for the combination of body and title fields

We now test our model's performance over structured document retrieval. A trivial way to process a structured document is to merge all fields into an unstructured one. This is used as our baseline for comparison. In addition to our approach, we also implemented two other approaches analyzed in Section 3.3.2: linear score combination (referred to as ScoreComb), and linear TF combination over Okapi (FreqComb).

Figure 5 shows the performance comparison over some tasks when combining body and title fields. In the experiments, the scores for the baseline, ScoreComb, and FreqComb are all generated by the BM25 formula (or, equivalently, GBM-Std)

having parameter $b=0.75$ (which is the optimal or suboptimal value for b in most tasks, except for 2003.td). For our approach, the parameter b is fixed as 0.75 accordingly. Given the fixed value of b , the baseline score is acquired (as the optimal score) by tuning parameter k_1 , while the scores for ScoreComb, FreqComb, and our approach in the figure are got by tuning k_1 (or $e(D)$ for our model) and the weights of fields.

From the figure, we can see that, our model is highly effective in dealing with structured documents.

5. CONCLUSION AND FUTURE WORK

In this paper, a gravitation-based IR model (GBM) was proposed, by adopting a physical perspective on information retrieval. In the basic version of GBM, each document is viewed as a bag of words and only TF, IDF, and document length is considered in modeling. The basic GBM model can not only give a *physical* interpretation of existing term weighting functions (e.g. BM25), but also derive new effective ranking formulas. The experimental results over some standard datasets and tasks show that the formulas derived from the basic GBM are among the most highly performance ranking functions proposed so far. In addition, as the derived term weighting functions satisfies all the heuristic constraints proposed in [5], they are more reasonable and hoped to be robust on various conditions. If document structure is considered, a novel approach for structured document retrieval can be naturally conducted from our model. The advantage of the approach is analyzed and tested by heuristics and experiments.

Most dominating IR models (including probabilistic models, language models, inference network models, etc) adopt a probabilistic perspective on information retrieval. Great successes have been and are being achieved with this perspective. However, we believe that viewing the problem from a different viewpoint is the same important as going deeper from traditional perspectives. This paper may be a first step to take a physical viewpoint. And we found that this new perspective gives us some insights for information retrieval.

As a totally new formal model, there are a lot of thing to do with it. First, as the derivation of the BM25 term weighting function from our model hints some internal (unknown) relationship between our model and other models, it may be interesting and necessary to study the relationship and possible combination between our model and existing models. Second, as pointed out in Section 3.4, this model has potential to include term proximity and static document ranking (which are extreme important in Web search) in its framework. Some work is needed to achieve this in the future. Finally, it is unclear whether and how the model can support some IR techniques, e.g. relevance feedback.

6. ACKNOWLEDGEMENTS

We would like to thank Guomao Xin, Ni Lao, and the anonymous reviewers for their helpful suggestions.

7. REFERENCES

- [1] G. Amati and C.J.V. Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transactions on Information Systems*, 20(4):357-389, 2002.
- [2] K. Sparck Jones and P. Willett, editors. *Readings in Information Retrieval*. Morgan Kaufmann, 1997.
- [3] G. Salton, A.Wong, and C.S. Yang. A vector space model for information retrieval. *Communications of the ACM*, 18(11): 613-620, Nov. 1975.
- [4] A. Singhal, C. Buckley, and M. Mitra. Pivoted document length normalization. In *Proceedings of SIGIR'96*.
- [5] H. Fang, T. Tao, and C. Zhai. A formal study of information retrieval heuristics. In *Proceedings of SIGIR'04*.
- [6] N. Fuhr. Probabilistic models in information retrieval. *The computer Journal*, Vol.35, No.3, pp 243-255.
- [7] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *SIGIR'01*, Sept 2001.
- [8] R. Baeza-Yates, and B. Ribeiro-Neto. *Modern Information Retrieval*, ACM Press, 1999.
- [9] S.E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of SIGIR'94*, 1994.
- [10] S.E. Robertson, C.J.V. Rijsbergen, and M.F. Porter. Probabilistic models of indexing and searching In *Proceedings of SIGIR'80*.
- [11] S. E. Robertson, S. Walker, and M. Beaulieu. Okapi at TREC-7: automatic ad hoc, filtering, VLC and filtering tracks. In *Proceedings of TREC'99*.
- [12] J. Ponte and W.B. Croft. A language modeling approach to information retrieval. In *Proceedings of SIGIR'98*.
- [13] F. Song and B. Croft. A general language model for information retrieval. In *Proceedings of SIGIR'99*.
- [14] S.K.M. Wong and Y.Y. Yao. On modeling information retrieval with probabilistic inference. *ACM Transactions on Information Systems*, 13(1), 69-99, 1995.
- [15] W.B. Croft. Combining approaches to information retrieval. In *Advances in Information Retrieval*, pp. 1-36. Kluwer, 2000.
- [16] S. Robertson, H. Zaragoza, and M. Yaylor. Simple BM25 extension to multiple weighted fields. In *Proceedings of CIKM'04*.
- [17] P. Ogilvie and J. Callan. Combining document representations for known item search. In *Proceedings of SIGIR'03*.
- [18] R. Wilkinson. Effective retrieval of structured documents. In *Proceedings of SIGIR'94*.
- [19] M. Lalmas. *Uniform representation of content and structure for structured document retrieval*. Technical report, Queen Mary and Westfield College, University of London, 2000.
- [20] S.H. Myaeng, D.H.Jang, M.S. Kim, and Z.C.Zhoo. A flexible model for retrieval of SGML documents. In *Proceedings of SIGIR'98*.
- [21] S. Shi, J.R. Wen, Q. Yu, R. Song, and W.Y. Ma. Gravitation-based model for information retrieval (extended version). Technique report, MSR-TR-2005-65, Microsoft Research, May 2005.
- [22] L. Page, S. Brin, R. Motwani, T. Winograd. *The PageRank Citation Ranking: Bringing Order to the Web*. Stanford Digital Libraries Working Paper, 1998.
- [23] TREC main page: <http://trec.nist.gov/>
- [24] B. Croft and J. Lafferty, editors. *Language Modeling for Information Retrieval*. Kluwer Academic Publishers, 2003.