Notes on IR

Nathan Chappell

March 30, 2020

Notes from [1]

- Probability Ranking Principle

  > The probability ranking principle (Robertson, 1977) states that for optimal performance, documents should be ranked in order of probability of relevance to the request or information need, as calculated from whatever evidence is available to the system

  This is a nice principle, but may not be appropriate for a system designed to help someone learn, rather than just retrieve information. Given context (i.e. user state), some information that is not "most relevant" to what the user has asked may be more advantageous from the perspective of learning. This is more in line with the notion of trying to maximize "information content" transmitted across a channel, i.e. some documents not most relevant to a query may help a user learn more, perhaps even if they are not as releva Given context (i.e. user state), some information that is not "most relevant" to what the user has asked may be more advantageous from the perspective of learning. This is more in line with the notion of trying to maximize "information content" transmitted across a channel. In particular, what a student must see in order to learn what they want to learn may be very indirectly related to a query, especially if the query is ill-founded (consider a wise teacher helping his students learn, he should be more of a guide than a presenter of information).

- RSJ Weight

  $$p_i = P(\text{document contains } t_i \mid \text{document is relevant})$$
  $$q_i = P(\text{document contains } t_i \mid \text{document is not relevant})$$

  $$RSJ(t_i) = \log \frac{p_i(1-q_i)}{q_i(1-p_i)}$$

  This can be given a simple estimate in terms of counting. Note that p_i and $(1-q_i)$ can be called "hits," in that the events correspond to a term's appearance and relevance coinciding. Let the complimentary events be

called "misses," and the function $-\log(P(\cdot))$ be the "measure of surprise" of an event (something like "information content"). Then RSJ is the difference between the measure of surpise of misses and hits. I.e. a term is weighted heavily if that term having a miss is very surprising, while getting a hit is not very surprising.

Note that, while it was Robertson's intention to remove the information-theoretic content from tf-idf, I'm putting it back into his measure.

- RSJ as IDF I don't think I'm smart enough to see how a fudged RSJ being equivalent to a naive Bayes model is better than the information-theoretic explanation given by Aizawa.

  He is definitely right that the question of event space is troubling... It is hard to see what the right space should be. My initial thought was to remove the notion of event space altogether, and consider the query/ response action as a act of communication – an attempt by multiple parties to compute a function using a channel as infrequently as possible.

- Document Normalization E.g., if a document is very long, under certain circumstances it should be normalized. It may be long because the author is wordy, or it may be long because the actual content is vast.

Notes from [2]

- Definition

  > Communication complexity studies ways to arrange *communications* between several parties so that at the end of the day the learn what they are supposed to learn, and to do this in the most efficient, or least *complex* way.

- Protocol Given inputs $x \in X$ (Alice), $y \in Y$ (Bob), we seek to calculate $f(x, y)$ transmitting as few bits as possible.

x				y
$f_1(x)$				
	$\rightarrow$	a_1	$\rightarrow$	
				$f_2(y, a_1)$
	$\leftarrow$	b_1	$\leftarrow$	
$f_1(x, a_1, b_1)$				
	$\rightarrow$	a_2	$\rightarrow$	
				$f_2(y, a_1, b_1, a_2)$
	$\leftarrow$	b_1	$\leftarrow$	
$\vdots$		$\vdots$		$\vdots$
	$\rightarrow$	a_t	$\rightarrow$	
				$f_2(y, \ldots, a_t) = f(x, y)$

The cost associated with the above protocol instance $P = P(f_1, f_2)$ is

$$C(P, x, y) = \sum_{i=1}^{t} |a_i| + |b_i|$$

There is a useful and general argument for lower bounding the complexity of a distributed problem made by considering the space of all possible message sequences (histories R_h) possibly made by a protocol $\{0,1\}^C(P)$, then relating these to coverings by combinatorial rectangles of the input space. The main idea is that the only thing that needs to be considered is $R_h = (a_1, b_1, \ldots, a_t, b_t)$. I.e. f_1 depends only on x, and $(b_1, b_2, \ldots)$ (this is some sort of an independence property). The argument is similar to the discrete geometric argument relating coverings to $\delta-$nets on a ball: we upper bound the size by volumetric and counting arguments. Here, we lower bound the complexity because the complexity must be "complex enough" to cover all the rectangles. The rectangles exists due to the "independence" mentioned before.

Notes from [3]

- Looks useful:

  > The core idea of Ingwersen and Jarvelin's framework is "how evidence of a searcher's information behavior may be applied to guide or adjust algorithmic information processing in system components through IR interaction"

- *Intellectual Perspective*
  - Information Searching
    interaction, goals/ tasks, behaviors/ strategies
  - Information Retrieval
    representing, storing, and finding information objects
  - Both
- *Theoretical Orientation*
  - People
  - Technology
  - Information

  Not all seventeen concepts

  Information – Information Searching
  - *Hierarchical Relationship of Information*
  - *Perceived Benefits*

Information – Both

- *Relevance*

Information – Information Retrieval

- *Representation* (sum of attributes)
- *Ranking*
- *Document Similarity*

People – Information Searching

- *Least Effort*
- *Iterative Process*

People – Both

- *Interaction*
- *Resolving an Uncertainty*

People – Information Retrieval

- *Provision* (assumed that providing information helps a user accomplish some task)

Technology – Information Searching

- *Channel Preference*

Technology – Both

- *Information Obtainability*

Technology – Information Retrieval

- *Query* (transformed input)
- *Neutrality of Technology*
- *Memex Vision*

Multiple Definitions: Information has physical (entropy), cognitive (state-change), and affective (touchy-feely) "aspects." As-process, as-knowledge, as-thing, as-attribute.

Hierarchical Relationships: DIKW: *data, information, knowledge, wisdom* (database, **IR**, **IS**, AI)

Information Ranking: In information searching, information ranking also is a notion well accepted; however, information searching researchers tend to focus more on the cognitive, affective, or contextual factors that determine the evaluation of search results and eventually the usefulness of information.

Relevance
: Saracevic(2007b) defined relevance as a relation between information and contexts (e.g., information need, intent, topic, problem) based on some property reflecting a relevance manifestation (e.g., topicality, utility, cognitive match).

Query
: a set of one or more symbols that is combined with other syntax and used as a command for an information retrieval systemto locate possibly relevant content indexed by that system

Memex Vison
: By nature though, both fields focus on the benefits of the use of technology, with little consideration of other options. As Rosenberg (1974) stated, the computer is not just a tool or machine but rather it is a way of looking at the world (p. 264) for these fields.

Information Obtainability
: information will be used in direct proportion to how easy it is to obtain (Summit, 1993).

  The lines of research in both fields hone in on making information easier to access, in terms of interfaces, expression of need or query, contextual help, and information visualization.

Notes from [4]

- A major assumption of the probability ranking principle (PRP) for IR is that

  > the relevance of a document to a query is independent of the relevance of other documents the user has seen before. The task addressed by the PRP is the user's scanning through the list of ranked documents.

- Often user's want different answers to the same problem (aspectual recall), and query reformulation is a more crucial (i.e. 'expensive') user operation.

- Model Assumptions:

  1. Only consider functional level of interaction
     e.g. don't consider the cost of an action depending on the interface used.
  2. Decisions are the major activity
     system offers binary choices to user. User may accept, in which case the decision is **positive**. If the user does not wish to modify the decision immediately after learning its consequences, the decision is **correct**.
  3. Users evaluate choices in linear order
  4. Only positive, correct decisions are of benefit to the user

- *Situations.* A *situation* consists of the list of choices, the user's information need, and the system's knowledge of the user. A positive decision changes the situation. The information need is static for a given situation.

- **Expected benefit to user**

  Event Space

  S set of situations

  s_i situation i

  C_i set of choices for s_i

  n_i $|C_i|$

  c_{ij} a choice for s_i

  Probabilistic Parameters

  p_{ij} probability c_{ij} is accepted in s_i

  q_{ij} probability choice is correct

  Cost Parameters

  e_{ij} effort of making choice c_{ij}

  b_{ij} benefit of making choice c_{ij}

  g_{ij} effort of correcting choice c_{ij}

  - The independence assumption is that p_{ij} is independent of choices *rejected* before (incorrect past choices).
  - The formula:

    $$E(c_{ij}) = e_{ij} + p_{ij}(q_{ij}b_{ij} + (1 - q_{ij})g_{ij})$$

    Average Benefit

    $$a_{ij} = q_{ij}b_{ij} + (1 - q_{ij})g_{ij}$$

    $$E(c_{ij}) = e_{ij} + p_{ij}a_{ij}$$

- Choices with $p_{ij} = 0$ are implicitly excluded from the list, as their expected benefit will be negative ($e_{ij} < 0$).

- Three strategies for maximizing c_{ij}:

  1. minimize e_{ij} (effort)
  2. maximize p_{ij} (selection probability)
  3. maximize g_{ij} (correctness probability)

- Expected benefit of list. Given a list of choices $r_i = \langle c_{i1}, c_{i2}, \ldots, c_{in} \rangle$, the expected benefit is

  $$r_i = e_{i1} + p_{i1}a_{i1} + (1 - p_{i1})(e_{i2} + p_{i2}a_{i2} + (1 - p_{i2})(e_{i3} + p_{i3}a_{i3} + \cdots$$

  $$r_i = \sum_{j=1}^{n} \left[(e_{ij} + p_{ij}a_{ij}) \prod_{k=1}^{j-1} (1 - p_{ik}) \right]$$

  Note the similarity of this to a Markov Model...

- **Comparing choice orders**
  Let

  $$t_i^{l,l+1} = (e_{il} + p_{il}a_{il}) \prod_{k=1}^{l-1} (1 - p_{ik}$$

  then we can compare the expected benefit of a list with the list made from transposing two elements by analyzing the term

  $$t_i^{l,l+1} - t_i^{l+1,l}$$

  If this term is non-negative, then the terms should not be transposed (c_{il} and $c_{i(l+1)}$ are in the right order in list r_i). After some algebra, the following ranking function is arrived at:

  $$\rho(c_{ij}) = a_{ij} + \frac{e_{ij}}{p_{ij}}$$

  Remember that e_{ij} is negative!

- **IIR-PRP** *(Probability ranking principle for interactive information retrieval)*
  Order the list of choices in decreasing values of

  $$\rho(c_{ij}) = a_{ij} + \frac{e_{ij}}{p_{ij}}$$

- Parameter Groups

  p_{ij} (Selection Probability)
  make dynamic for changing information requirements

  e_{ij}, g_{ij}, q_{ij} (Effort Parameters)
  UI, expertise of user

  a_{ij} (Benefit)
  general problem, entropy

Notes from [5]

- *Relevance Clues* Different users use similar criteria, but apply different weights depending on the user, class of user, task, and progress in task.

  Criteria:
  Information Factors

  - Content
    topic, quality, depth, scope, currency, treatment, clarity
  - Object
    type, organization, representation, format, availability, accessibility, costs

  - Validity
    accuracy, authority, verifiability
    Human Factors
  - Situational Match
    usability, urgency, value
  - Cognitive Match
    understanding, *novelty*, mental effort
  - Affective Match
    emotional response, fun, frustration, uncertainty
  - Belief Match confidence, personal credence

- Relevance feedback improves user performance, but users tend not to use it!

- *Central assumptions in IR evaluation* (Weak) *Relevance is:*
  - Topical
  - Binary
  - Independent
  - Stable
  - Consistent

- *No Consistency* Relevancy judgements can be very inconsistent among judges (expect no better than 70% among experts, closer to 30/50% among different groups).

- "different retrieval systems are better at finding the highlyrelevant documents than those that are better at finding generally relevant documents."

- > Subject expertise seems to be one variable that accounts strongly for differences in relevance inferences by group of judges – higher expertise results in higher agreement, less differences.

- *Relevancy Measures* It seems that interval measures, e.g. $[0, 1]$ are better than binary (relevant/ not-relevant) and graded (due to bias). The measures tend to be bimodal, with peaks at the ends note to self: maybe something like a beta distribution? Look into "subjective logic," it seems that an application of trust networks and subjective bayesian logic in IR may be novel and useful

- *Independence*

  > Information objects presented early have a higher proba-bility of being inferred as relevant... Relevance judg-ments do change as information is added.

- *Conclusion* In the conclusion to the article Saracevic implies that there is a lot of room for research and, even more so, implementation of better relevance mechanisms in IR. It would seem that there is a lack.

Notes from [6]

- *Mutual Information*
  Considering a noisy channel, the question is how to measure the information received. If the channel is completely noisy, i.e. you get a 1 or 0 with $p = .5$ regardless of the transmission, this should correspond to receiving 0 information. Shannon continues:

  > Evidently the proper correction to apply to the amount of information transmitted is the amount of this information which is missing in the received signal, or alternatively the uncertainty when we have received a signal of what was actually sent. ... Following this idea the rate of actual transmission, R, would be obtained by subtracting from the rate of production (i.e., the entropy of the source) the average rate of conditional entropy.
  >
  > $$R = H(x) - H(x|y)$$

  Note the relationship between the rate R of conditional entropy, and Aizawas interpretation of tf-idf. I have a hunch that it may be useful to consider an IR, IIR, or IS-IR system as a communication model. The source of information would be the user and their uncertainty, the transponder would be the means of querying the system, and the noise is the user's inability to express their information need to the system (or, equivalently, the system's inability to understand the user's intent, whichever makes you feel better). The assumption would be that, were the user able to properly express their needs, the IR system would be able to provide the necessary information. tf-idf then corresponds to the method of presenting the choices with the highest mutual entropy in linear order. The **IIR-PRP** model is somewhat more elaborate, and seems to take into account *rate* more so than just **PRP**, by considering effort and probability of correctness.

Notes from [7]

- In this paper the question being addressed is how to approximate a *Relevance Model* without any training data.
- *Two approaches to determining relevance.*
  1. "Forward error:" Take a query, map it to documents, and snap it to the nearest existing data.
  2. "Backward error:" take the documents, model queries that should hit them, and snap queries to the best fitting models.
- *A Formal Relevance Model*

  > We use the term *relevance model* to refer to a mechanism that determines the probability $P(w|R)$ of observing a word w in the documents relevant to a particular information need. ... **Queries and relevant documents are random samples from an underlying relevance model R.**

This model R exists for every "information need."

"Recall $P(w|R)$ is the relative frequency with which we expect to see the word w during repeated independent random sampling of words from all of the relevant documents."

- *The basic formulae*
  Suppose we are given a sequence $q_1, \ldots, q_k$ of query words, being generated by some model R. What is the probability that the next word is w? This is the attempt to guess the model given a query.

$$P(w|R) \approx \frac{P(w, q_1, \ldots, q_k)}{P(q_1, \ldots, q_k)}$$

They propose two techniques for estimating the joint distribution P. Say we have a set $\mathcal{M}$ of models (distributions), and pick M with probability $P(M)$, then

$$\begin{aligned} P(w, q_1, \ldots, q_k) &= \sum_{M \in \mathcal{M}} P(M) P(w, q_1, \ldots, q_k | M) \\ &= \sum_{M \in \mathcal{M}} P(M) P(w|M) \prod_{i=1}^{k} P(q_i|M) \end{aligned}$$

This uses strong independence properties, the next technique assumes only independence between w and the q_i:

$$\begin{aligned} P(w, q_1, \ldots, q_k) &= P(w) \prod_{i=1}^{k} P(q_i|w) \\ &= P(w) \prod_{i=1}^{k} \sum_{M_i \in \mathcal{M}} P(M_i|w) P(q_i|w) \end{aligned}$$

Notes from [8]

- I didn't go through this article in excrutiating detail, I mostly just wanted to get the giss of "subjective logic." There's a manuscript that can give a better foundation if this seems like it will be a useful tool, but the idea is to allow for uncertainty in probabilistic opinions. For binary opinions there is a 1-1 mapping with the β-distribution.

- *Relevance.* This paper is the only one that I've seen that gives a reasonable, general definition of relevance in terms of probability theory.

$$\Psi(x|Y) = |p(x|y) - p(x|\bar{y})|$$

This is actually a nice definition, and has an interesting connection with mutual entropy and insightful derivation (this is my work, I didn't see it in the paper).

$$\begin{aligned} p(x) &= p(x,y) + p(x,\bar{y}) \\ &= p(x|y)p(y) + p(x|\bar{y})p(\bar{y}) \\ &= p(x|y)p(y) + p(x|\bar{y})(1-p(y)) \\ &= p(y)(p(x|y) - p(x|\bar{y})) + p(x|\bar{y}) \Rightarrow \\ |p(x) - p(x|\bar{y})| &= p(y)\Psi(x|Y) \end{aligned}$$

A derivation that I've made, although I'm not sure if it's useful yet, follows. If we let $R(x,y) = p(x|y) - p(x|\bar{y})$ (something like the "signed relevance"), a few manipulations give:

$$p(x) = \left(\frac{p(x|y) + p(x|\bar{y})}{2}\right) + \left(\frac{p(y) - p(\bar{y})}{2}\right) R(x,y)$$

This may be more enlightening if we can generalize it to random variables...

Anyways, the main point is that relevance is a measure of how "not independent" two entities are, weighted by some factor. Going from a notion of relevance to information seems challenging, and will considered in the future.

References

[1] *Understanding Inverse Document Frequency: on Theoretical Arguments for IDF*
Robertson, Stephen
Microsoft Research
Journal of Documentation 60

[2] *Communication Complexity*
Razborov, Alexander
University of Chicago
Steklov Mathematical Institue Moskow

[3] *The Seventeen Theoretical Constructs of Information Searching and Information Retrieval*
Jansen, Bernard
Pennsylvania State University
Rieh, Soo
University of Michigan
Journal of the American Society for Information Science and Technology

[4] *A Probability Ranking Principle for Interactive Information Retrieval*
Fuhr, Norbert
University of Duisberg–Essen, Germany
Information Retrieval, Springer 2008

[5] *Behavior and effects of relevance*
Tefko Saracevic
Rutgers University, NJ
Wiley InterScience 2007

[6] *A Mathematical Theory of Communication*
Shannon, Claude
Bell Systems Technical Journal 27
October, 1948

[7] *Relevance-Based Language Models*
Lavrenko, Victor; Croft, Bruce
Center for Intelligent Information Retrieval
University of Massachusetts
SIGIR'01 September 2001

[8] *Generalising Bayes' Theorem in Subjective Logic*
Josang, Audun
University of Oslo
IEEE MFI Conference
Open Campus, US Army Research Labs