Outline for an information theoretic search engine

It is proposed an information theoretic search engine is like RADAR. The query words are the emitted signals and the document database is the object to be detected. Various echoes come off the database, and analogous with echo cancelation, the signal with the lowest entropy is selected. Commensurate with Shannon's theory, low entropy documents are signal, higher entropy documents are noise. Thus, my proposal separates signal from noise. As many relevant documents can be tined to be signal as desired.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1400v3 | CC-BY 4.0 Open Access | rec: 7 Oct 2015, publ: 7 Oct 2015

Jacobus Vanderwilt

Information retrieval as a source channel process

Bird's eye view of Claude Shannon's work on the transmission of information.

Claude E Shannon proposed language is a key part of a communication process. This process models instances of communication as the transmission of information. A message is encoded by the transmitter, sent across a channel, where it has to contend with noise or irrelevant information, and decoded by the receiver. Imagine a couple of boys playing with a walkie talkie. One boy makes a statement in English. The transmitting walkie-talkie encodes this message into electromagnetic waves – these waves correspond in crucial ways to the original sound waves: in terms of volume and frequency of the boy's voice. But using electromagnetic waves introduces static or noise: bits of signal that are NOT the first boy's message. The null case of FM transmission is 'white noise': possibly moving particles generate radiowaves at all frequencies. Electromagnetic noise can interfere with the sent message to the extent this message becomes incomprehensible. In that case, noise supersedes message, and the message and the noise will have to be separated. This is the decoding process.

The bird's eye view of Shannon's communication theory, then, is: a message is encoded by a transmitter, sent across a noisy channel, and then decoded by the receiver. This is known as the channel source approach or noisy channel approach. Variants of this approach are at the base of assigning a correct Part of Speech out of many options to all words in a text, speech recognition, satellite communication, and Interatlantic telephone communication.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1400v3 | CC-BY 4.0 Open Access | rec: 7 Oct 2015, publ: 7 Oct 2015

Google and other search engines as instances of information transmission: a message (encoded as a probability vector) or query sent through a channel with noise (all documents in the database) to be decoded upon arrival at the receiver: distill relevant from irrelevant documents.

I will argue the process of information retrieval 'submitting key words or bits of language to a search engine and receiving a set of documents as a consequence' as a instance of Shannon's communication theory is revealing and advantageous, creating a new opportunity to experiment with search engines; possibly better search engines. This idea is not entirely new and has been proposed by Djoerd Hiemstra and Franciska de Jong, for example. See also http://www.cs.cmu.edu/~aberger/start/perspective.html

which comes close to my idea but is not quite the same.

The message in the information theoretic view of the information retrieval process is the user query. This message is now to be encoded, and I will describe a way of doing so. On the way through the noisy channel, the encoded message travels through much noise: all the documents in the database, both relevant documents and immaterial documents. The receiver is responsible for decoding signal from noise, which, in our case, means documents relevant to the query from immaterial documents.

The Information Theoretic Search Engine: the console of an IR system as RADAR.

Radar works as follows. A transmitter emits a signal - radiowaves. The signal is then bounced off the object the radar seeks to detect, for instance, an enemy ship. But the signal that is bounced back is not focused and is diffuse in the sense light coming off a candle is diffuse. So the receiver part of RADAR gets an echo: the waves that come off the enemy ship - several copies of that ship, much like several copies of your voice come back to you from an echo well.

While this is somewhat of a conjecture on my part, confirmed by using Google, 'echo cancelation' is an EE technique, used for picking out the right signal from all the echoes.

I propose a search engine/information retrieval system is like RADAR. The user emits a signal - the expressed information need or search request. The request then hits the database/document collection, and is returned to the user. A system can be set up – see below – where several answer sets are returned. The various answer sets are

analogous to the various copies of your voice bounced back from the echo well or the various copies of the ship.

Now to pick the copy that most clearly is your voice or a copy of the enemy ship. We do this with 'entropy minimization' of echo signals. We pick the copy (of your voice or the enemy ship) that is clearest, and discard the other copies as noise. We do this with Shannon's information theory and choose the signal with the lowest entropy. Shannon's theory, after all, is about separating signal from noise. Entropy minimization is a well-known technique for RADAR (in particular ISAR).

(Shannon's decoding of info sent across a noisy channel may be but a variant of Viterbi decoding. And this may be known in certain circles - just not in one I belong to. Both statistical models pick the right signal from a number of signals or signal+noise.

Googling "low entropy" and "viterbi path" yields an answer set. I cannot find a DIRECT answer to my question about the equivalency between Viterbi's decoding and Shannon decoding. But my working hypothesis is that they are closely related and possibly the same: select the message with the lowest entropy. The article at the following link seems to confirm this: http://www.ists.dartmouth.edu/library/281.pdf)

Key concepts in Shannon's work: every day concepts and the mathematificiation of these.

I will discuss my understanding of some key concepts in the theory of Shannon, followed by the exposition in a primer by William Benish, who uses Shannon's 'information theory', as it is known, for medical diagnostics. I will elaborate how and why below.

Key concepts in Shannon's theory are key concepts in language, the vehicle for communication. There is, for instance, redundancy. Language has much redundant information. Leaving out every second letter of the alphabet in a sentence, does not render this sentence incomprehensible. Word-initial 'l' in English is never followed by an 's'. Reading an 'l', specification the next symbol is not an 's' is certain and can thus be predicted. Redundancy, predictability and uncertainty are thus narrowly related, and all three are features of written as well as spoken language. The mathematical tool for redundancy, predictability, and uncertainty is probability theory. Probability theory is essential to information theory.

Connected to redundancy, predictability, and certainty – the flip side of uncertainty - is the notion of order or 'lack of chaos'. The greater the certainty about the next symbol

given one symbol, the greater the lack of chaos or order of a message in a language. Shannon formalized the interconnection between all these notions.

As an example, let's discuss the notions 'improbable' (and thus uncertain) and 'degree of surprise'. The less probable an event, the greater someone's surprise at it occurring. Stealing a jocular reflection on this from William Benish, the probability of a nice sunny day in the Netherlands being low, surprise at its occurrence will be commensurately high. Shannon formalized this notion by defining 'surprisal'. Surprisal is the negative of the logarithm of a probability distribution:

$S(p)=-\log_2 p$.

If an impossible event occurs, no large enough number exists to express the surprisal associated with it.

Similarly, Shannon's work consists of the mathematical expression of the interconnectedness of every day notions. We presuppose familiarity on the reader's part of probability theory, of the theory of logarithms, and will expose Shannon's formalization of 'chaos' – which he terms 'entropy' -, information.

I will then explain how one way of doing information retrieval is already an execution of my idea in practice, although to my knowledge it is has not been designated as such. After this, I will present pseudo-code for my proposal. Finally, I will discuss how my idea is an answer to the so-called 'long tail' problem in information retrieval.

Benish' exposition of Shannon's work.

Benish (2000) provides us with a refresher of logarithms, introduces surprisal, and discusses some facets of probability theory. Suppose we have a probability distribution over diseases, with p(heart pain)=0.5, p(gastroesephageal reflux) =.25, p(chest wall pain)=.125, ad p(some other disease)=.125. A patient has one of these conditions, and the total of the sum of these probability distributions is therefore 1. At this point a guessing game is introduced. If we want to guess which disease a patient really has, we'd guess the disease with the largest probability distribution. This guess will be right half the time and wrong half the time. If we are wrong, we would then choose chest wall pain, and we would be right ¼ of the time. If we were wrong, we'd pick one of the remaining possibilities, and would be right ¼ of the time. The number of guesses we would need is (1 x 1/2) + (2 x 1/4) + (3 x 1/4) = 7/4. This is called the *expected value* of the number of guesses – Benish explains the concept *expected value* earlier in his exposition. The expected value of the *surprisals* therefore is: $E[S(p)] = -1/2 \log_2(1/2) - 1/4 \log_2(1/4) - 1/8 \log_2(1/8) - 1/8 \log_2(1/8) = -1/2(-1) - 1/4(-2) - 1/8(-3) - 1/8(-3) = 7/4$. Benish then states: "Entropy is the expected value of the surprisal. It is the amount, on

average, that we will be surprised when we learn the truth.". If entropy is low, order must be high and we are not surprised at what we find. The general formula for entropy is:"H(X) = -SUMMATIONi=1 to n p(xi)log(pxi).

Encoding a search request to a probability vector.

Now suppose we have a search request and suppose it consists of a string of words. We can compile a probability vector by calculating p(w|d) for each document.

for each document
  for each word
   calculate p(w1|d1, p(w2|d1), p(w3|d1) .....A
             p(w2|d2), p(w2|d2), p(w3|d2) .....B

A and B (and C and D etc) are probability vectors, a series of numbers. This has been the encoding process: each word in the query has yielded a probability vector. If a word does not occur in a document, p(w|d)=0. Otherwise, each probability vector, that is, each word, will indicate a set of numbers. Calculate the entropy for each word's term/document product. In my proposal, relevant answers are the signal and irrelevant answers are the noise. Use Shannon's entropy formula to determine which term/document pair has the lowest entropy. That document set is the signal; other document sets contain more noise. Determine MAX(H(X), H(Y)) where H(X) and H(Y) are two instances of the word/document pair in the probability vectors. From a CS assignment instantiating the idea a message is selected with the lowest entropy: "The program should find a decoded message with the lowest entropy. The message with the lowest entropy is our best guess for the original text. Note: We will assume the frequencies are the same" http://www.mathcs.emory.edu/~dsavenk/courses/fall13/cs170/hw/hw6.pdf.

We could construct bigram language models rather than unigram models as above by following the explanation in Salesky. The query would then consist of bigrams and the entropy number associated with the probability vector.

If we picked bigrams constructed of words on which the additional demand be imposed they belong to a grammatical category such as Subject, Object, Verb, (Grammatical feature selection, see van der Wilt 2015) performance might improve even further.

Example.

A query word that does not occur very frequently has probability .17. A word occurring more frequently in another document may be .73. A partial probability vector may be (.17 .73). Applying Shannon entropy to such a partial probability vector will yield:

(.17*log2(.17)+.73* log2(.73))
(.17.-2.55+)+.73-.454)

The small probability .17 is compensated with a larger surprisal, -2.55. This means the contribution of each search term to the total entropy is evened out. Words that are not very frequent yet contribute to diminishing the entropy of an answer set (d1.d4,d7).

The above proposal is not as innovative as it may seem at first. The Kullback Leibler Divergence has *also* been proposed to perform document retrieval. The lower the relative entropy of a document to the standard of the query, the more likely the document is to be ranked higher in the answer set a query returns. Here, too, the message with the lowest entropy is returned. The encoded message are the terms in the relative entropy formula derived from the query; the formula terms derived from each single document in the database determine whether that document is signal or noise.

Channel source and the long tail issue

On https://ciir.cs.umass.edu/research/longqueries/ Bruce Croft writes: "Long queries represent a small but significant percentage of the queries submitted to web search engines currently. In other applications, such as collaborative question answering where people ask questions for other people to answer, long queries are typical, rather than unusual. Many information needs can be more easily expressed using longer, sentence-length queries, but the inadequacies of current search engines force people to try to think up the right combination of keywords to find relevant documents. This can be very difficult and often leads to search failures. On the other hand, long queries are handled poorly by current search engines. This is due at least in part to these queries being part of the "long tail", meaning that they are infrequent and lack many of the statistical features that are used for effective ranking of short queries. Being able to effectively handle long queries would represent a significant advance in the capability of search engines from the user's point of view, and should substantially improve our understanding of the underlying information retrieval process."

I suggest that assigning to each word in each document a probability distribution for as many search words as you like in the manner I describe above, i.e. by picking the answer set with the lowest entropy, considering all documents in the database as noise and signal and determining signal, might be a suitable way to enable long queries.

The RADAR model and related information-theoretic approaches.

The model presented here and other attempt to view information retrieval as instances of the channel source model may be related. One such attempt is proposed by Bergeron (2003). Bergeron claims documents in the answer set that are relevant

represent signal, and documents that are irrelevant are noise. As low entropy is related to a high signal to noise ratio, Bergeron's approach may be related to the RADAR model. Some more formal work would be required to work this out in greater detail.

In general, traditional approaches to Information Retrieval such as TF*IDF and Signal to Noise retrieval may be related. It seems we have a family of related approaches when we view information retrieval through the prism of information theory. One effort relating TF*IDF to information theory is Wong and Yao (1992). The authors establish that " both the IDF and S/N weighting schemes can be considered as approximations of the proposed information theoretic measure for term specificity."

Please refer to the section *encoding a search request as a probability vector*. The probability vectors there are: e.g. (w1|d1, w2|d1, w3|d1….), (w1|d2, w2|d2, w3|d2.). Let's make a small change and incorporate the occurrence of w1, w2…wn in the entire document collection: W1, W2, W3…… The vectors are then: (w1/W1|d1, w2/W2|d1, w3/W3|d). The average info for all documents is

SUMMATION-i p(w-k/W-k)*logp(w-k/W-k), where i is all documents.

This summation incorporates TF and IDF and is probably equivalent to it. The summation is also known as Average Information, or entropy of term w-k in the collection, of which Noise is the negative - see http://facweb.cs.depaul.edu/mobasher/classes/csc575/lectures/lecture3.pptx. Now find the term/document pair, or the top term/document pairs and these are the ideal answers to the query.

References

Benish, William, The Application of Information Theory to Diagnostic Testing: A Primer, http://filer.case.edu/wab4/public/primer2.pdf

Bergeron, Brian P, Bioinformatics Computing, Pearson Publications, 2003

http://facweb.cs.depaul.edu/mobasher/classes/csc575/lectures/lecture3.pptx

Gavish,Matan A Crash Course on Shannon's Mutual Information for Categorical Data Analysys.

http://web.stanford.edu/~gavish/documents/information_in_cat_data_analysis.pdf

Fisher, David Laguage Modeling and Information Retrieval.

http://sourceforge.net/p/lemur/wiki/Language%20Modeling%20and%20Information%20Retrieval%20Background/

http://www.cs.cmu.edu/~aberger/start/perspective.html

Salesky, Liz. Math 20, Project Paper, N-gram Language Models

Wong and Yao, An Information-Theoretic Measure of Term Specifity, Dept of Computer Science, University of Regina, Regina, Sasketchewan, Canada,