Conditional Probability and Information Retrieval

Huafeng Zhang

I. Introduction

When searching references in a database, people will use different keywords and operators. But often, they don't know which keywords work best to find what they want. Therefore, software engineers write algorithms to rank associated phrases based on the keywords. By ranking the phrases, software engineers help people figure out the most related key terms for their search. We will use Dey and Majumdar' s's paper about fast mining of phrases to explain how they use conditional probability to solve a particular problem in the field of information retrieval.

The searches discussed here are more like a Google advanced search than like a normal Google search. These searches are conducted on information databases such as newspaper databases or scholarly article collections. When searching on this kind of database, you will use operators like *AND, OR,* and *NOT* with keywords to find relevant references.

In 1979, Van Rijsbergen introduced the Probability Ranking Principle [2](*PRP*) in this way:

> *If a reference retrieval system's response to each request is a ranking of the documents in the collection in order of decreasing probability of relevance to the user who submitted the request, where the probabilities are estimated as accurately as possible on the basis of whatever data have been made available to the system for this purpose, the overall effectiveness of the system to its user will be the best that is obtainable on the basis of those data.*[2]

Even though Van Rijsbergen doesn't mention conditional probability here, when he talks about ranking documents "in order of decreasing probability of relevance to the user," we understand that he is using conditional probability to define "relevance." Just as he used conditional probability to formulate a principle for ranking documents, similarly, Dey and Majumdar use conditional probability to build algorithms for ranking "interesting" phrases.

Interesting phrases, as they define the term, appear in the documents that turn up in response to a query (D') but occur less frequently in the entire database (D). These phrases can make the selected documents distinctive from the rest of the database, which can help people find the most relevant keywords for their references.

II. Conditional Probability

How does conditional probability help us define what is interesting (or distinctive) about a subset of a database? In order to tackle this question, we should know what conditional probability is. It is the probability that one event happens given another event has occurred. Even though non-statisticians don't know the term conditional probability, they use the concept frequently to make decisions and judgements. For example, they will be less likely to lend money to people who borrowed their money but haven't returned it, and they tend to bring umbrellas when they see the cloud getting dark etc.

For statisticians, conditional probability is a strictly defined term. Let's suppose that there are events A and B in the sample space S and P(B) >0 , then the conditional probability of A given B can be written as $P(A \mid B)=\frac{P(A \cap B)}{P(B)}$, where P(A ∩ B) is the probability that both events A and B

occur. P(A | B) is always between zero and one, and the larger P(A | B) is, the more likely it is that event A will happen, given event B[3].

For statisticians, conditional probability is everywhere, from a tiny p value and a significance level α to complicated model designing. P value[4] is the probability of obtaining results at least as extreme as those observed, given that the null hypothesis is true, i.e. P value = P(extreme observations | null hypothesis). The statistical significance level[5], α, is the probability of rejecting the null hypothesis, given that it is true , i.e. α = P(reject null hypothesis | null hypothesis). When building models, statisticians need to decide if it is appropriate to drop some coefficients given p value and significance level. If the p value of the coefficient is relatively larger than the statistical significance level α, they will consider dropping the coefficients to fit data with simpler model.

III. How information retrieval uses conditional probability

In Dey and Majumdar' s study[6] of finding interesting phrases to differentiate a subset of documents from the whole database, they assume that for phrases that have "high interestingness for the query,"... "the occurrence of the keywords... are conditionally independent of each other." That is to say, there are n mutually independent keywords q1, q2...qn. in the query Q for any phrase p that has high interestingness.

They model a scoring function "where the score of a phrase p is computed as the probability of occurrence of the phrase p in the chosen sub-collection D' normalized by its probability of occurrence in the entire corpus." i.e. $S_D(p, Q) = \frac{Pd'(p)}{P(p)}$, where Q={q1, q2, …, qn}. The numerator Pd'(p) is P(p | [{q1,q2,...,qn}, O]), where O is short for "Operator" including the *AND, NOT* and

OR operators. To rewrite P(p | [{q1,q2,...,qn}, O]), they use Bayes' theorem[7], $P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$ because they know P(qi | p) and P(p) instead of P(p | qi) for any keyword in the query Q.

After applying Bayes' theorem, Dey and Majumdar get $P(p \mid [\{q1,q2,...,qn\}, O]) = \frac{P([\{q1,q2,...,qn\}, O] \mid p) \times P(p)}{P(p \mid [\{q1,q2,...,qn\}, O])}$. When scoring all the phrases given one selected query, then P(p | [{q1,q2,...,qn}, O])=1 (because Q is the sample space[8] that has collectively exhaustive keywords q1, q2, q3 ···, that is, one of the keywords must occur, and the union of the keywords cover all the keywords in the query Q). So $P(p \mid [\{q1,q2,...,qn\}, O]) \approx P([\{q1,q2,...,qn\}, O] \mid p) \times P(p)$. Finally, after substituting $P(p \mid [\{q1,q2,...,qn\}, O]) = P([\{q1,q2,...,qn\}, O] \mid p) \times P(p)$ into their phrase scoring function $S_D(p, Q) = \frac{Pd'(p)}{P(p)}$, they find $S_D(p, Q) = P([\{q1,q2,...,qn\}, O] \mid p)$.

But how to calculate P([{q1,q2,...,qn}, O] | p)? In order to do this, we need to use the Rule of Multiplication [9] and know how to compute the conditional probability for two conditional independent events.

Sample space A is a set of events that include partitional events A1, A2, A3···An, then P(A1A2A3···An) = P(A1)P(A2 | A1)P(A3 | A1A2)···P(An | A1A2···An-1) for $n \geq 2$[9] In this example, the sample space is the set of queries Q={q1,q2,...qn}. So the probability of finding results with all of the queries q1, q2,...qn is P(q1,q2,...qn) ($n \geq 2$)

$$P((q1 \mid p)(q2 \mid p)...(qn \mid p)) = P(q1 \mid p)P([q2 \mid p] \mid [q1 \mid p])P([q3 \mid p] \mid [(q1 \mid p)(q2 \mid p) \; P([qn \mid p] \mid [(q1|p)\,(q2|p)\;\; (q(n-1)|p)])$$

When events A and B are independent from each other, then the probability that both events A and B happen equals the product of the probability for event A happen and event B happen, i.e. $P(A \cap B)=P(A)P(B)$[10] . After substituting $P(A \cap B)=P(A)P(B)$ into conditional probability formula $P(A \mid B)=\frac{P(A \cap B)}{P(B)}$, we find $P(A \mid B)=\frac{P(A \cap B)}{P(B)}=\frac{P(A)P(B)}{P(B)}=P(A)$. Similarly, in Dey and Majumdar' s study, the keywords q1, q2,...qn in the query Q for any given phrase p are mutually independent (the assumption that was discussed before), that is,

$$P([q2 \mid p] \mid (q1 \mid p))=P(q2 \mid p),$$

$$P([q3 \mid p] \mid [(q1 \mid p)(q2 \mid p)])=P(q3 \mid p),$$

...

$$P([qn \mid p] \mid [(q1 \mid p)(q2 \mid p)\ \ (q(n\text{-}1) \mid p)])=P(qn \mid p).$$

So the conditional probability of getting results in a database with mutually independent terms q1,q2,...qn for given phrase p by using the *AND* operator is $P(q1, q2,...qn \mid p, AND)$:

$$P(q1, q2,...qn \mid p, AND) = \frac{P(q1....q(n-1)qn), p)}{P(p)}$$

$$= \frac{P(q1....q(n-1)qn \mid p)P(p)}{P(p)}$$

$$= \frac{P(q1 \mid p)....P(q(n-1) \mid p)P(qn \mid p)P(p)}{P(p)}$$

$$= P(q1 \mid p)...P(q(n-1) \mid p)P(qn \mid p)$$

$$i.e.,\ P((\bigcap_{i=1}^{n} qi) \mid p) = \prod_{i=1}^{n} P(qi \mid p)$$

Therefore, for a query Q=[p, {q1,q2,...qn}, *AND*], $P((\bigcap_{i=1}^{n} qi) \mid p) = \prod_{i=1}^{n} P(qi \mid p)$ which is used as the Dey and Majumdar' s phrase scoring method[6].

Using the same assumption, Dey and Majumdar derive the phrase scoring function for use with the *OR* operator. The process is close to the phrase scoring method used with the *AND* operator. In both, they apply conditional probability to derive the phrase scoring function.

$$P[p,\{q1,q2,...qn\},OR\ ]=P(q1 \mid p)+P(q2 \mid p)+...+P(qn \mid p)-P(q1,q2 \mid p)-...+(-1)^{n-1}P(q1,q2,...qn \mid p)$$

$$\text{i.e.}^{[6]},\ P[p,\{q1,q2,...qn\},OR\ ]=\sum_{i=1}^{n} P(qi \mid p)-\sum_{i\neq jx\in(i,j)} \prod P(qx \mid p)+\ ...\ +(-1)^{n-1}\prod_{i=1}^{k} P(qi \mid p)$$

They model the phrase scoring functions for listing interesting phrases used with the *AND* and *OR* operator by using conditional probability, then they build the algorithms and test the experiment quality. Explaining the algorithms and experiment is beyond the scope of the paper, so I will not discuss these two sections.

IV. Conclusion

How to find interesting phrases for given queries is a problem in conditional probability because in order to solve the problem, software engineers need to know the probability of the occurrence of phrase p given query Q first. Then, they can build the algorithms to list the interesting phrases in the database. As discussed in this paper, conditional probability is used to help software engineers do fast mining of interesting phrases in a database in an effort to provide their users better information retrieval services.

Reference

[1] Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*.

[2] Blair, D. C. (1979). Information Retrieval, 2nd ed. C.J. Van Rijsbergen. London: Butterworths; 1979: 208 pp. Price: $32.50. *Journal of the American Society for Information Science, 30*(6), 113-114. doi:10.1002/asi.4630300621

[3] Conditional probability. (2016). Retrieved November 13, 2016, from https://en.wikipedia.org/wiki/Conditional_probability

[4] P-value. (2016). Retrieved November 13, 2016, from https://en.wikipedia.org/wiki/P-value

[5] Statistical significance. (2016). Retrieved November 14, 2016, from https://en.wikipedia.org/wiki/Statistical_significance

[6] Dey, A., & Majumdar, D. (2014). Fast Mining of Interesting Phrases from Subsets of Text Corpora. Retrieved from https://pdfs.semanticscholar.org/9e58/a347ddf4762a1f6ee56b825223c304be2231.pdf.

[7] Bayes' theorem. (2016). Retrieved November 17, 2016, from https://en.wikipedia.org/wiki/Bayes'_theorem

[8] Sample space. (2016). Retrieved November 19, 2016, from https://en.wikipedia.org/wiki/Sample_space

[9] Conditional Probability. (2016). Retrieved December 03, 2016, from http://www.math.uah.edu/stat/prob/Conditional.html

[10] Conditional Probability and Independent Events. (n.d.). Retrieved November 19, 2016, from http://www.cut-the-knot.org/Curriculum/Probability/ConditionalProbability.shtml