User Models & Metrics

Georges Dupret

August 6, 2009

Summary

1. What are the common assumptions about user behavior implicit or explicit in common metrics?
2. We identify essentially two classes:
   - Assume the user effort is fixed and estimate the session success,
   - Assume the session is successful and estimate the effort.
3. We argue that:
   - Metrics parameters can be estimated thanks to the associated user model,
   - It would be better to fix neither utility nor effort (Pareto frontier),
   - Instead of comparing metrics, we should compare user models.

Part 1

Utility Based Metrics

Discounted Cumulated Gain at rank R

DCG Metric[1]

$$DCG_r = \sum_1^R \frac{1}{D_r} rel_r$$

where D_r is a discounting factor increasing with the rank r.

Motivation

- *Utilitarian:* The utility of a document to a user decreases when the document is low in the ranking.
- *Probabilistic:* All documents are not examined with the same probability. Search Engine logs suggest that the probability of seeing a document follows a power law in r.

[1] [Järvelin and Kekäläinen, 2002]

DCG at rank R: User Model

- Before examining the result list the user chooses a number between 1 and R according to the probability

$$p(r^*) = \frac{D_{r^*}}{\sum_{r=1}^{r=R} D_r}$$

- The session utility is measured by the amount of seen relevance.

The expected session utility is:

$$☺ = \sum_r^R p(r) rel_r \quad = \quad \sum_r^R \frac{D_{r^*}}{\sum_{r=1}^{r=R} D_r} rel_r = \frac{DCG_R}{\sum_{r=1}^{r=R} D_r} \propto DCG_R$$

DCG at rank R: User Model (cont.)

Observations

1. There is no distinction between one session with n clicks and n sessions with one click; Users have no memory.
2. The search need not be sequential.
3. User search effort is constant; The metric is related to the amount of seen relevance only.

DCG: User Model Estimation[2]

Events:

A user clicks ($c = 1$) on a document d at position r if

1. he examines the position ($e_r = 1$),
2. the document is attractive enough to grant a click ($a_r = 1$).

Probabilities:

1. The probability of examination depends only on the position,
2. A document is attractive if it is relevant (There is no difference between perceived and actual relevance).

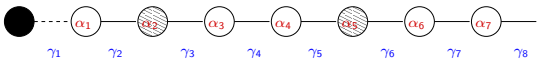
[2] [Dupret and Piwowarski, 2008, Guo et al., 2009]

Discounted Cumulated Gain: Estimation

The probability of a click is:

$$P(c = 1|r, d) \quad = \quad P(e = 1|r) \times P(a_r = 1) = \gamma_r \alpha_d$$

For example, the likelihood of:

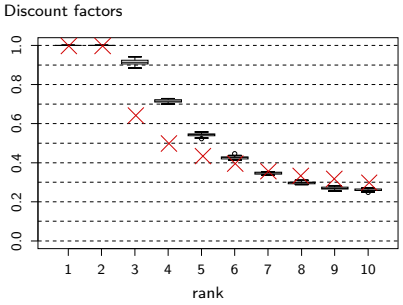


is

$$L = (1 - \gamma_1 \alpha_1) \times \gamma_2 \alpha_2 \times (1 - \gamma_3 \alpha_3) \ldots$$

1. We multiply the likelihood of all sessions,
2. We maximize with respect to the γ_r and α_d
3. We obtain estimates for $\gamma_r = P(e = 1|r)$ and $\alpha = P(a = 1|rel_d)$.

DCG: Discounting Factors γ_r[3]



[3] [Dupret et al., 2007]

Discounted Cumulated Gain: Metric

In all generality, we are interested in the join distribution

$$P(\text{utility, session; ranking})$$

A natural metric in this context is the expected success:

$$\begin{aligned} ☺ &= \mathbb{E}(\text{utility; ranking}) \\ &= \sum_r P(c_r|e_r)P(e_r) = \sum_r \alpha_r \gamma_r \propto DCG_R \end{aligned}$$

DCG: Related Metrics & Observations

Related Metrics

- robust DCG
- Ranked Biased Precision in [Moffat and Zobel, 2008]
- idea: Instead of one trial, say the user makes a pair of trials, each at a distinct rank, with $P(r_1, r_2)$.

Observations:

- Making the model explicit helped identifying shortcomings and suggests improvements.
- We can use observational data to estimate the parameters of the model.
- The metric can be expressed as a marginalization of the user model distribution.

Part 2

Effort Based Metric

Mean Average Precision

The average of the precisions at the relevant documents.

$$MAP = \frac{1}{R}\sum_{r=1}^{\infty} \text{precision at } r \times \text{ relevance at } r$$

User Model

- The user decides how many *relevant* documents he needs –say k– and browses sequentially until he finds them [Robertson, 2008].
- [Moffat and Zobel, 2008]: "Every time a relevant document is encountered, the user pauses, asks "Over the documents I have seen so far, on average how satisfied am I" and writes a number on a piece of paper. Finally, when the user has examined every document in the collection –because this is the only way to be sure that all of the relevant ones have been seen– the user computes the average of the values they have written."

Mean Average Precision (cont.)

Relation between the user model and the metric.

1. The level of a user happiness is the precision at k.
   - amount of relevance needed to achieve success is fixed.
   - precision is related to the effort.
2. We don't know the proportion of users who want exactly k documents, hence we assume a uniform distribution.

MAP: Estimation

1. The user want $k = 1$ relevant document and $r_2 = 0$,
2. The user want $k = 2$ documents and both $r_2 = r_5 = 1$.
3. The user want $k > 2$ documents but there are only 2 attractive documents in the collection.

$$\begin{aligned} L &= P(k=1)P(a_1=0)P(a_2=1,r_2=0) \\ &\quad \times P(a_3=0)P(a_4=0)P(a_5=1,r_5=1) \\ &+ P(k=2)P(a_1=0)P(a_2=1,r_2=1) \\ &\quad \times P(a_3=0)P(a_4=0)P(a_5=1,r_5=1) \\ &+ (1-P(k=1)-P(k=2))P(a_1=0)P(a_2=1) \\ &\quad \times P(a_3=0)P(a_4=0)P(a_5=1)\prod_{i>5} P(a_i=0) \end{aligned}$$

MAP: Metric

Diagnostic Metric : Compute $\mathbb{E}$(precision) based on the observations,

Predictive Metric : For the same ranking as above and $k = 1$, if we know $r_2 = 1$ and $r_5 = 1$, we have three possible sessions (ℓ is the search length):

1. d_2 is clicked; $\ell = 2$ (probability $P(a_2 = 1, r_2 = 1)$),
2. d_2 is skipped and d_5 is clicked; $\ell = 5$ (probability $P(a_2 = 0, r_2 = 1) \times P(a_5 = 1, r_5 = 1)$),
3. both d_2 and d_5 are skipped; $\ell = L$ (probability $P(a_2 = 0, r_2 = 1) \times P(a_5 = 0, r_5 = 1)$)

MAP: Metric (cont.)

The MAP user model provides the join distribution:

$$P(\ell, \text{session}; \text{ranking})$$

A "natural" measure for the effort is the search length ℓ but MAP uses instead the precision as a reward:

$$\text{☺} \quad = \quad \mathbb{E}(\text{precision}) = \sum_k \sum_{\text{session}(k)} \frac{k}{\ell} P(\ell, \text{session(k)}; \text{ranking})$$

MAP: Improvements

- Relax the uniform distribution assumption on k:
  - parametrized distribution over k.
  - estimates from click-through logs.
- It also suggests other improvements. For example: for a fixed k, use the number R_k of retrieved relevant documents while searching for k relevant documents to compute a new "precision":

$$\text{☺} \quad = \quad \sum_k^{\infty} \frac{R_k}{\ell_k} P(k) rel_k$$

where ℓ_k is the search length.

In the Same Family

Reciprocal Rank: The reciprocal rank of the first relevant document. The user model is the MAP user model with $P(k = 1) = 1$.

bpref family: [Buckley and Voorhees, 2004, Sakai, 2007] By summing over the number of relevant documents, the model implicitly divide users according to the number of documents they need, like the MAP user model. How the *effort* is estimated varies among the different versions.

Expected Search Length: [Brookes, 1968, Bollmann and Raghavan, 1988] the expected number of documents the user must read before finding a desired number of relevant documents (Cooper, 1968).

Cascade Model [Craswell et al., 2008, Chapelle and Zhang, 2009]

Part 3

Where to go from here?

Utility & Effort

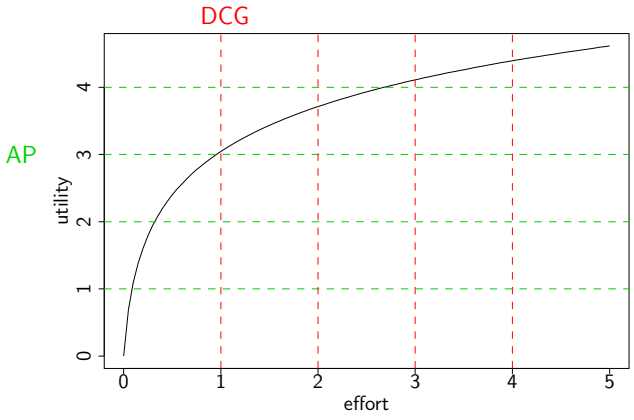
Two classes of metrics:

- DCG fix the effort and marginalize over the utility, MAP fix the utility and marginalize the effort.
- The two metrics are related to the marginalization over the utility / effort

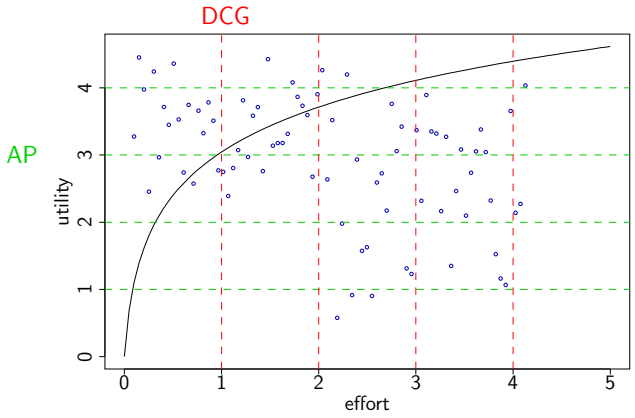
1. User models incorporate both utility and effort to predict session success,
2. A metric derived from such a user model scales naturally: If we know $P(\text{success}, \text{utility}, \text{effort}, \text{session}|\text{ranking function})$ then

$$\text{☺} = \mathbb{E}(\text{success}|\text{utility}, \text{effort}, \text{ranking function})$$

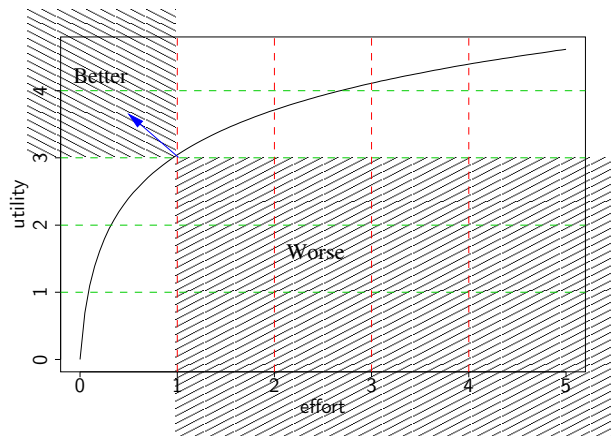
Utility & Effort: One Dimensional Metrics



Utility & Effort: Optimizing One Metric



Utility & Effort: Comparing Ranking Function



Utility & Effort: Conclusions

1. We need a metric that includes both effort & utility,
2. This metric needs a realistic user model,
3. The best user model is the one with the best predictive power,
4. The join probability offers a scale free method to compare models

$$P(\text{success}_1 > \text{success}_2, \text{utility}, \text{effort})$$

User Models

- Beware of models... navigational queries are very frequent...
- User choices during a search are limited; We can take advantage of the imposed structure to model user behavior.
  - Example of using the structure: [Piwowarski et al., 2009, Piwowarski et al., 2007],
  - Metric proposal relying on user making choices and decisions: [Fuhr, 2008].

Getting clicks: *www.historyse.com*

Conclusions

1. Common metrics either fix utility and measure effort, or fix effort and compare utilities; We should develop metrics that model both.
2. Utility and effort represent a trade-off; Sometimes model comparison makes no sense.
3. Metrics are hard to compare or evaluate, but they can be matched with a marginalization on the distribution of their respective user model.
   - Marginalizing resolve the scaling problem between effort and utility,
   - We should base metrics comparison on the associated user model.

Thanks!

- To the organizing commitee,
- To Justin Zobel and Alistair Moffat for suggestions.

Bollmann, P. and Raghavan, V. V. (1988).
A utility-theoretic analysis of expected search length.
In *SIGIR '88: Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 245–256, New York, NY, USA. ACM.

Brookes, B. C. (1968).
The measure of information retrieval effectiveness proposed by swets.
Journal of Documentation, 24:41–54.

Buckley, C. and Voorhees, E. M. (2004).
Retrieval evaluation with incomplete information.
In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 25–32, New York, NY, USA. ACM.

Chapelle, O. and Zhang, Y. (2009).
A dynamic bayesian network click model for web search ranking.
In *WWW '09: Proceedings of the 18th international conference on World wide web*, pages 1–10, New York, NY, USA. ACM.

Craswell, N., Zoeter, O., Taylor, M., and Ramsey, B. (2008).
An experimental comparison of click position-bias models.
In *First ACM International Conference on Web Search and Data Mining WSDM 2008*.

Dupret, G., Murdock, V., and Piwowarski, B. (2007).
Web search engine evaluation using clickthrough data and a user model.
In *WWW2007 workshop Query Log Analysis: Social and Technological Challenges*.

Dupret, G. and Piwowarski, B. (2008).
A user browsing model to predict search engine click data from past observations.
In Press, A., editor, *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*.

Fuhr, N. (2008).
A probability ranking principle for interactive information retrieval.
Information Retrieval, Springer.

Guo, F., Liu, C., and Wang, Y. M. (2009).
Efficient multiple-click models in web search.
In *WSDM '09: Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 124–131, New York, NY, USA. ACM.

Järvelin, K. and Kekäläinen, J. (2002).
Cumulated gain-based evaluation of ir techniques.
ACM Transactions on Information Systems (ACM TOIS), 20(4):222–246.

Kelly, D. (2009).
Methods for Evaluating Interactive Information Retrieval Systems with Users", *http:/dx.doi.org/10.1561/1500000012*, volume 3 of *Foundations and Trends in Information Retrieval*.

Moffat, A. and Zobel, J. (2008).
Rank-biased precision for measurement of retrieval effectiveness.
ACM Trans. Inf. Syst., 27(1):1–27.

Piwowarski, B., Dupret, G., and Jones, R. (2009).
Mining user web search activity with layered bayesian networks or how to capture a click in its context.
In *WSDM '09: Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 162–171, New York, NY, USA. ACM.

Piwowarski, B., Gallinari, P., and Dupret, G. (2007).
An extension of precision-recall with user modelling (PRUM): Application to XML retrieval.
Transactions on Information Systems (TOIS).

Robertson, S. (2008).
A new interpretation of average precision.
In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 689–690, New York, NY, USA. ACM.

Sakai, T. (2007).
Alternatives to bpref.
In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 71–78, New York, NY, USA. ACM.

Voorhees, E. M. and Harman, D., editors (2005).
TREC: Experiment and Evaluation in Information Retrieval.
MIT press.

Yilmaz, E., Aslam, J. A., and Robertson, S. (2008).
A new rank correlation coefficient for information retrieval.
In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 587–594, New York, NY, USA. ACM.