IJCSC International Journal of Computer Science & Communication Vol. 1, No. 2, July-December 2010, pp. 137-139

Modified Discounted Cumulative Gain – a New Measure for Subgraphs

E.R.Naganathan[1], S. Narayanan[2] & K.Ramesh Kumar[3]

[1]Department of Computer Applications, Velammal College of Management and Computer Studies, Chennai.
[2,3]Dept. of Computer Sci. & Engg., Alagappa University, Karaikudi-630002. Tamil Nadu.

ABSTRACT

This paper proposes a novel approach to apply the concept of "Lift" into Discounted Cumulative Gain of Subgraph mining algorithms. The application of lift in subgraph mining can be treated as Modified Discounted Cumulative Gain (MDCG). In this paper, we show that the best finding of MDCG according to lift metric. We also show how this concept can be generalized to find MDCG. This MDCG may be applied to the various applications instead of DCG.

1. INTRODUCTION

"Lift" is the most commonly used metric to measure the performance of targeting models. The purpose of targeting model is to identify a subgroup from a larger population. Generally, lift can be calculated by looking at the cumulative targets captured up to p% as a percentage of all targets and dividing by p% ie., Lift is simply the ratio of target response divided by average response.[1].

In this paper we present a computationally efficient algorithm to find the Discounted Cumulative Gain (DCG) using Lift. This modified DCG using lift may plays an efficient role in other DCG applications.

2. DISCOUNTED CUMULATIVE GAIN (DCG)

Discounted Cumulative Gain (DCG) is a measure of effectiveness of a web search engine algorithm or related applications often used in information retrieval [2]. Using a graded relevance scale of documents in a search engine result set, DCG measures the usefulness, or gain, of a document based on its position in the result list. The DCG is given by

$$DCG_p = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{\log_2(1+i)}$$

3. MODIFIED DISCOUNTED CUMULATIVE GAIN (MDCG)

3.1. Lift

Lift is the statistical definition of dependence of two sets A and B which is given by

$$\frac{P[A \cap B]}{P[A]P[B]}$$

with the obvious extensions to more than two sets [3].

Lift originally called Interest, was first introduced by Motwani, et al., (1997), it measures the number of times X and Y occur together compared to the expected number of times if they were statistically independent.[4]

The Lift can also be framed as a function of confidence [5]

$$\text{lift}(A \to C) = \frac{|D|\,\text{conf}(A \to C)}{\sup(C)}$$

Where

Support of a graph is given by [6]

In a given graph F_G, the support F_S^G is defined as

$$Sup(F_G) = F_S^G = \frac{\text{number of graph transactions F}}{\text{total number of graph transactions}}$$

And confidence is given by [6]

Given two induced subgraph F_b and F_h, the confidence of the association rule

$F_b \Rightarrow F_h$ is defined as

$$Conf\,(F_b \Rightarrow F_h) = \frac{\text{number of graphs } F \text{ where } F_b \cup F_h \subset F \in FD}{\text{number of graph } F \text{ where } F_b \subset F \in FD}$$

By this conviction, Lift is obviously monotone in confidence and unaffected by rule support when confidence is held fixed.

$$Lift = \left(F_{Lift}^G\right) = \frac{\text{number of graph of } F_b \text{ and } F_h}{\text{number of } F_b \; x \text{ number of } F_h} = \frac{P(XUY)}{P(X)\;P(Y)}$$

In the case of subgraph architecture, lift can be defined as

$$Lift = \left(F_{Lift}^G\right) = \frac{\text{number of graph of } F_b \text{ and } F_h}{\text{number of } F_b \; x \text{ number of } F_h} = \frac{P(XUY)}{P(X)\;P(Y)}$$

The relationship of X and Y are defined by the lift as:

i) lift value > 1 then X and Y depend on each other

ii) lift value < 1 then X depends on the absence of Y or vice-versa

iii) lift value close to 1 then X and Y are independent.

3.2. Modified Discounted Cumulative Gain (MDCG)

The premise of DCG is that highly relevant documents appearing lower in a search result list should be penalized as the graded relevance value is reduced logarithmically proportional to the position of the result. Here, the new measure called 'lift" is applied to the DCG. Then the Modified Discounted Cumulative Gain accumulated at a particular rank position p is defined as:

$$MDCG_p = lift(F_1) + \sum_{i=2}^{p} \frac{lift(F_i)}{\log_2 i}$$

There has not been shown any theoretically sound justification for using a logarithmic reduction factor [7] other than the fact that it produces a smooth reduction. An alternative formulation of MDCG places much stronger emphasis on retrieving relevant documents ranked higher using a power distribution and is formulated as :

$$MDCG_n = \sum_{i=1}^{n} \frac{2\,\mathrm{lift}(F_i)-1}{\log_2(1+i)}$$

3.3. Algorithm for Modified Discounted Cumulative Gain

Input : Graphs F_1, F_2, …. $F_{k\text{-}2}$

Output : Modified Discounted Cumulative Gain

Step 1 : Construct the rules for the given sub graph

Step 2 : Calculate the Lift value for each subgraph such that

The relationship of X and Y are defined by the lift through the condition

i) if lift value > 1 then X and Y depend on each other

ii) if lift value < 1 then X depends on the absence of Y or vice-versa

iii) if lift value close to 1 then X and Y are independent.

Step 3: Then calculate the MDCG value such as

$$MDCG_p = \mathrm{lift}(F_1) + \sum_{i=2}^{p} \frac{\mathrm{lift}(F_i)}{\log_2 i}$$

Construction of Rule

In a given graph F_G, the support F_S^G is defined as

$$Sup(F_G) = F_S^G = \frac{\text{number of graph transactions F}}{\text{total number of graph transactions}}$$

Given two induced subgraph F_b and F_h, the confidence of the association rule

$F_b \Rightarrow F_h$ is defined as

$$\mathrm{Conf}(F_b \Rightarrow F_h) = \frac{\text{number of graphs } F \text{ where } F_b \cup F_h \subset F \in FD}{\text{number of graph } F \text{ where } F_b \subset F \in FD}$$

If the value of sup(F_G) is more than a threshold value minsup, F_G is called as frequent induced subgraph.

3.4. Example for the Algorithm

Consider the following graph data set.

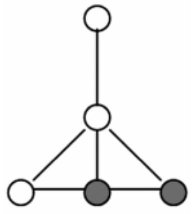

Fig. 1

The number of frequent subgraphs are :

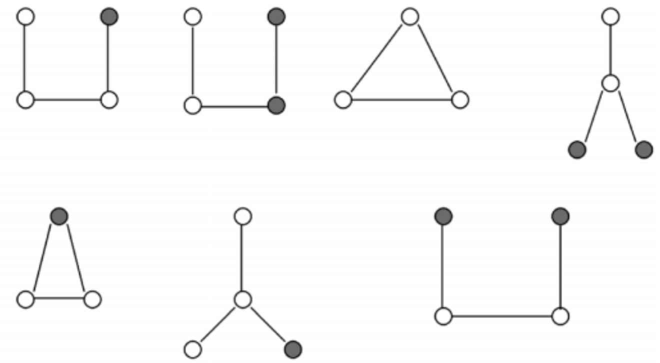

Fig. 2

Here the subgraphs listed are F_1, F_2, …. F_7 [8]

Some sample "lift" values are calculated according to the subgraphs in Fig – 2. The P(X) and P(Y) are taken for six rules that is labeled as $R1$, $R2$ …. $R6$.

Rules	*P(X)*		*P(Y)*
R1	W → W → W	⟶	R
R2	W → W	⟶	R → R
R3	R → W	⟶	W → R
R4	W → W	⟶	R
R5	W → W	⟶	W
R6	R → W	⟶	W

MDCG Table

i	*Rule*	*lift*	*Log i*	*Lift/Log i*	*MDCG*
1	R1	0.037037037	0	0	0.037037037
2	R2	0.090909091	1	0.090909091	0.127946128
3	R3	0.027777778	1.584962501	0.017525826	0.145471954
4	R4	0.070707071	2	0.035353535	0.18082549
5	R5	0.016042781	2.321928095	0.00690925	0.187734739
6	R6	0.045751634	2.584962501	0.017699148	0.205433887

4. RESULTS AND DISCUSSION

This sample was from the data set of chemical compound and synthetic data set. The complete summary of result are as follows :

Here seven frequent subgraphs are considered with the threshold value of σ = 66%. The rule for frequent subgraph are then applied to find out the new measure 'lift' then the value of lift at each position of the rule is taken into the calculation to find out the MDCG. All MDCG calculations are then relative value on the interval 0.0 to 1.0 and so are cross-query comparable. Hence the MDCG value at each step may be efficient in other applications instead of DCG.

REFERENCES

[1] Why Lift? – Data Modeling and Mining – Information Management Online, June 21, 2002.

[2] Discounted Cumulative Gain – Wikipedia, the Free Encyclopedia.

[3] Sergey Brin, R.Motwani and C.Silverstein – Beyond Market Baskets : Generalizing Association Rules to Correlations – SIGMOD'97 AI, USA.

[4] Sergey Brin, Rajeev Motwani, Jeffrey D. Ullman, and Shalom Tsur, "Dynamic Itemset Counting and Implication Rules for Market Basket Data", *In SIGMOD 1997, Proceedings ACM SIGMOD International Conference on Management of Data*, pages 255-264, Tucson, Arizona, USA, May 1997.

[5] R.J.Bayardo Jr and Rakesh Agrawal – Mining the Most Interesting Rules – Procedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 145-154, 1999.

[6] An Apriori-based Algorithm for Mining Frequent Substructures from Graph Data, Akihiro Inokuchi, Takashi Washio and Hiroshi Motoda, PKDD2000, Sept. 13-16, 2000, Lyon, France.

[7] B.Croft, D.Metzler and T.Strohman (2009) Search Engines. Information Retrieval in Practice. Adison Wesley.

[8] Michihiro Kuramochi and George Karypis, *Frequent Subgraph Discovery IEEE International Conference on Data Mining*, (ICDM), 2001.