IR Evaluation Methods for Retrieving Highly Relevant Documents

What is the Problem ?

Traditional IR Evaluation Methods categorizes documents as **relevant** or **irrelevant**. With these two values we cannot evaluate the system by its ability to retrieve most relevant documents.

What is Proposed ?

This paper proposes evaluation methods which uses **multiple relevance levels** so that we can evaluate IR system better.

For instance, relevance **level 0** for not relevant documents. Relevance **level 1** for marginally relevant documents, relevance **level 2** for fairly relevant documents and **level 3** for mostly relevant documents.

Two methods are proposed which are :

- P – R Curves **for each relevance level** with **separate recall base** and
- Cumulative gain and Discounted cumulative gain vectors.

We are using **3 relevance level**. (Level 0 is ignored since it is irrelevant)

We are using **3 query types** SUM, SSYN-C, WYSN. We are also using **query expansion** on these query types. Therefore, total of **6 query types** are used in case study.

P – R Curves

We are using classic P – R curves for each relevance level with separate recall base.

This way we can see the difference in performance better.

6 query types are used.

Cumulative Gain and Discounted Cumulative Gain

Highly relevant documents are more important than marginally relevant documents.

Thus, **as we increase the rank**, naturally the **documents become less important**.

We can do it by **adding their relevance level to each other** and calculating a Cumulative Gain. In this way, if we retrieve highly relevant documents, we add more to cumulative gain which means next document becomes less important.

However, **rank 1** highly relevant document and **rank 100** highly relevant document **contributes same to this cumulative gain**. To prevent that, we use **discounting function** such as log function to reduce the effect as the rank increases.

Case Study

- Mostly relevant documents really benefit from strongly typed queries
- Expansion on strongly typed queries increase the precision value in P – R curves
- Strongly typed queries with expansion are twice more effective than other query types.
- As we increase the relevance level, CG and DCG value differences between queries becomes proportionally larger.