5. Novelty & Diversity
5.1. Why Novelty & Diversity?
5.2. Probability Ranking Principled Revisited
5.3. Implicit Diversification
5.4. Explicit Diversification
5.5. Evaluating Novelty & Diversity
1. Why Novelty & Diversity?

- Redundancy in returned results (e.g., near duplicates) has a negative effect on retrieval effectiveness (i.e., user happiness).

- No benefit in showing relevant yet redundant results to the user.

- Bernstein and Zobel [2] identify near duplicates in TREC GOV2; mean MAP dropped by 20.2% when treating them as irrelevant and increased by 16.0% when omitting them from results.

- Novelty: How well do returned results avoid redundancy?
1. Why Novelty & Diversity?

- **Redundancy in returned results** (e.g., near duplicates) has a **negative effect** on retrieval effectiveness (i.e., user happiness)

- **No benefit** in showing **relevant yet redundant** results to the user

- Bernstein and Zobel [2] identify **near duplicates** in TREC GOV2; mean **MAP dropped by 20.2%** when treating them as **irrelevant** and **increased by 16.0%** when **omitting them** from results

- **Novelty**: How well do returned results avoid redundancy?
Why Novelty & Diversity?

- **Ambiguity of query** needs to be reflected in the returned results to account for uncertainty about the user’s information need.

  ![Jaguar](image1.png) ![Question Mark](image2.png) ![Jaguar](image3.png) ![Eclipse](image4.png) ![Defender](image5.png) ![Cookies](image6.png)

- **Query ambiguity** comes in different forms:
  - **topic** (e.g., jaguar, eclipse, defender, cookies)
  - **intent** (e.g., java 8 – download (transactional), features (informational))
  - **time** (e.g., olympic games – 2012, 2014, 2016)

- **Diversity**: How well do returned results reflect query ambiguity?
Implicit vs. Explicit Diversification

- **Implicit diversification methods** do not represent query aspects explicitly and instead operate directly on document contents and their (dis)similarity
  - Maximum Marginal Relevance [3]
  - BIR [11]

- **Explicit diversification methods** represent query aspects explicitly (e.g., as categories, subqueries, or key phrases) and consider which query aspects individual documents relate to
  - IA-Diversify [1]
  - xQuad [10]
  - PM [7,8]
2. Probability Ranking Principle Revisited

If an IR system’s response to each query is a ranking of documents in order of decreasing probability of relevance, the overall effectiveness of the system to its user will be maximized.

(Robertson [6] from Cooper)

- Probability ranking principle as bedrock of Information Retrieval

- Robertson [9] proves that ranking by decreasing probability of relevance optimizes (expected) recall and precision@k under two assumptions
  - probability of relevance \( P[R|d,q] \) can be determined accurately
  - probabilities of relevance are pairwise independent
Probability Ranking Principle Revisited

- Probability ranking principle (PRP) and the underlying assumptions have shaped retrieval models and effectiveness measures.
  - **retrieval scores** (e.g., cosine similarity, query likelihood, probability of relevance) are determined looking at documents in isolation.
  - **effectiveness measures** (e.g., precision, nDCG) look at documents in isolation when considering their relevance to the query.
  - **relevance assessments** are typically collected (e.g., by benchmark initiatives like TREC) by looking at **(query, document) pairs**.
3. Implicit Diversification

- Implicit diversification methods do not represent query aspects explicitly and instead operate directly on document contents and their (dis)similarity
3.1. Maximum Marginal Relevance

Carbonell and Goldstein [3] return the next document $d$ as the one having maximum marginal relevance (MMR) given the set $S$ of already-returned documents.

$$\arg \max_{d \notin S} \left( \lambda \cdot \text{sim}(q, d) - (1 - \lambda) \cdot \max_{d' \in S} \text{sim}(d', d) \right)$$

with $\lambda$ as a tunable parameter controlling relevance vs. novelty and $\text{sim}$ a similarity measure (e.g., cosine similarity) between queries and documents.
3.2. Beyond Independent Relevance

- Zhai et al. [11] generalize the ideas behind Maximum Marginal Relevance and devise an approach based on language models.

- Given a query $q$, and already-returned documents $d_1, \ldots, d_{i-1}$, determine next document $d_i$ as the one minimizes

$$\text{value}_R(\theta_i; \theta_q)(1 - \rho - \text{value}_N(\theta_i; \theta_1, \ldots, \theta_{i-1}))$$

- with $\text{value}_R$ as a measure of relevance to the query (e.g., the likelihood of generating the query $q$ from $\theta_i$),

- $\text{value}_N$ as a measure of novelty relative to documents $d_1, \ldots, d_{i-1}$,

- and $\rho \geq 1$ as a tunable parameter trading off relevance vs. novelty.
The novelty value $N$ of $d_i$ relative to documents $d_1, \ldots, d_{i-1}$ is estimated based on a two-component mixture model

- let $\theta_O$ be a language model estimated from documents $d_1, \ldots, d_{i-1}$
- let $\theta_B$ be a background language model estimated from the collection

the log-likelihood of generating $d_i$ from a mixture of the two is

$$l(\lambda|d_i) = \sum_v \log((1 - \lambda) P[v | \theta_O] + \lambda P[v | \theta_B])$$

the parameter value $\lambda$ that maximizes the log-likelihood can be interpreted as a measure of how novel document $d_i$ is and can be determined using expectation maximization
4. Explicit Diversification

- **Explicit diversification methods** represent query aspects explicitly (e.g., as categories, subqueries, or topic terms) and consider which query aspects individual documents relate to.

- **Redundancy-based explicit diversification methods** (IA-SELECT and xQUAD) aim at covering all query aspects by including at least one relevant result for each of them and penalizing redundancy.

- **Proportionality-based explicit diversification methods** (PM-1/2) aim at a result that represents query aspects according to their popularity by promoting proportionality.
4.1. Intent-Aware Selection

- Agrawal et al. [1] model **query aspects as categories** (e.g., from a topic taxonomy such as the Open Directory Project)
  - query \( q \) belongs to category \( c \) with probability \( P[c|q] \)
  - document \( d \) relevant to query \( q \) and category \( c \) with probability \( P[d|q,c] \)

- Given a query \( q \), a baseline retrieval result \( R \), their objective is to find a set of documents \( S \) of size \( k \) that maximizes

\[
P[S \mid q] := \sum_c P[c \mid q] \left( 1 - \prod_{d \in S} (1 - P[d \mid q, c]) \right)
\]

which corresponds to the **probability that an average user finds at least one relevant result** among the documents in \( S \)
**Intent-Aware Selection**

- Probability $P[c|q]$ can be estimated using **query classification methods** (e.g., Naïve Bayes on pseudo-relevant documents)

- Probability $P[d|q,c]$ can be decomposed into
  - probability $P[c|d]$ that document belongs to category $c$
  - query likelihood $P[q|d]$ that document $d$ generates query $q$

**Theorem:** Finding the set $S$ of size $k$ that maximizes

$$P[S|q] := \sum_c P[c|q] \left( 1 - \prod_{d \in S} (1 - P[q|d] \cdot P[c|d]) \right)$$

is **NP-hard** in the general case (reduction from MAX COVERAGE)
**IA-SELECT (Greedy Algorithm)**

- **Greedy algorithm** (IA-SELECT) iteratively builds up the set $S$ by selecting document with **highest marginal utility**

\[
\sum_c P[\neg c | S] \cdot P[q | d] \cdot P[c | d]
\]

with $P[\neg c | S]$ as the probability that none of the documents already in $S$ is relevant to query $q$ and category $c$

\[
P[\neg c | S] = \prod_{d \in S} (1 - P[q | d] \cdot P[c | d])
\]

which is initialized as $P[c | q]$
Submodularity & Approximation

- **Definition**: Given a finite ground set \( N \), a function \( f: 2^N \rightarrow \mathbb{R} \) is **submodular** if and only if for all sets \( S, T \subseteq N \) such that \( S \subseteq T \), and \( d \in N \setminus T \), \( f(S \cup \{d\}) - f(S) \geq f(T \cup \{d\}) - f(T) \)

- **Theorem**: \( P[S|q] \) is a submodular function

- **Theorem**: For a submodular function \( f \), let \( S^* \) be the optimal set of \( k \) elements that maximizes \( f \). Let \( S' \) be the \( k \)-element set constructed by greedily selecting element one at a time that gives the largest marginal increase to \( f \), then \( f(S') \geq (1 - 1/e) f(S^*) \)

- **Corollary**: IA-SELECT is \((1-1/e)\)-approximation algorithm
4.2. eXplicit Query Aspect Diversification

- Santos et al. [10] use query suggestions from a web search engine as query aspects.

- **Greedy algorithm**, inspired by IA-SELECT, iteratively builds up a set $S$ of size $k$ by selecting document having highest probability

\[
(1 - \lambda) \cdot P[d | q] + \lambda \cdot P[d, \neg S | q]
\]

where $P[d|q]$ is the document likelihood and captures **relevance** and $P[d, \neg S|q]$ is the probability that $d$ covers a query aspect not yet covered by documents in $S$ and captures **diversity**.
Probability $P[d, \neg S | q]$ can be decomposed into

$$\sum_i P[\neg S | q_i] \cdot P[q_i | q]$$

Probability $P[q_i | q]$ of subquery (suggestion) given query $q$ estimated as uniform or proportional to result sizes

Probability $P[\neg S | q_i]$ that none of the documents already in $S$ satisfies the query aspect $q_i$ estimated as

$$P[\neg S | q_i] = \prod_{d \in S} (1 - P[d | q_i])$$
IA-SELECT and xQUAD Criticized

- Redundancy-based methods (IA-SELECT and xQUAD) degenerate
  - IA-SELECT does not select more results for a query aspect, once it has been fully satisfied by a single highly relevant result, which is not effective for informational intents that require more than one result
  - IA-SELECT starts selecting random results, once all query aspects have been satisfied by highly relevant results
  - xQUAD selects results only according to \( P[d|q] \), once all query aspects have been satisfied by highly relevant results, thus ignoring diversity
4.3. Diversity by Proportionality

- Dang and Croft [7,8] develop the proportionality-based explicit diversification methods PM-1 and PM-2

- Given a query $q$ and a baseline retrieval result $R$, their objective is to find a set of documents $S$ of size $k$, so that $S$ proportionally represents the query aspects $q_i$

- **Example**: Query *jaguar* refers to query aspect *car* with 75% probability and to query aspect *cat* with 25% probability

  $$S_1 = \{d_1, d_2, d_3, d_4\} \quad S_2 = \{d_1, d_2, d_5, d_6\} \quad S_3 = \{d_1, d_2, d_5, d_7\}$$

  $S_1$ more proportional than $S_2$ more proportional than $S_3$
Sainte-Laguë Method

- **Ensuring proportionality** is a classic problem that also arises when **assigning parliament seats** to parties after an election.

- **Sainte-Laguë method** for seat allocation as used in New Zealand:
  - Let $v_i$ denote the number of votes **received** by party $p_i$.
  - Let $s_i$ denote the number of seats **allocated** to party $p_i$.
  - While not all seats have been allocated:
    - assign next seat to party $p_i$ with highest quotient $\frac{v_i}{2s_i + 1}$.
    - increment number of seats $s_i$ allocated to party $p_i$.

<table>
<thead>
<tr>
<th>Party</th>
<th>Votes</th>
<th>Seats</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>48%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>22%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>16%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>14%</td>
<td></td>
</tr>
</tbody>
</table>
PM-1

- PM-1 is a naïve adaption of the Sainte-Laguë method to the problem of selecting documents from $D$ for the result set $S$

- Members of parliament (MoPs) belong to a single party only, hence a document $d$ represents only a single aspect $q_i$, namely the one for which it has the highest probability $P[d|q_i]$

- Allocate the $k$ seats available to the query aspects (parties) according to their popularity $P[q_i|q]$ using the Sainte-Laguë method

- When allocated a seat, the query aspect (party) $q_i$ assigns it to the document (MoP) $d$ having highest $P[d|q_i]$ which is not yet in $S$

- Problem: Documents relate to more than a single query aspect in practice, but the Sainte-Laguë method cannot handle this
PM-2

PM-2 is a **probabilistic adaption** of the Sainte-Laguë method that considers to what extent documents relate to query aspects.

- Let \( v_i = P[q_i|q] \) and \( s_i \) denote the proportion of seats assigned to \( q_i \).
- While not all seats have been allocated:
  - **select query aspect** \( q_i \) with highest quotient
    \[
    \frac{v_i}{2s_i + 1}
    \]
  - **select document** \( d \) having the highest score
    \[
    \lambda \cdot \frac{v_i}{2s_i + 1} \cdot P[d|q_i] + (1 - \lambda) \cdot \sum_{j \neq i} \frac{v_j}{2s_j + 1} \cdot P[d|q_j]
    \]
    with parameter \( \lambda \) trading off relatedness to aspect \( q_i \) vs. all other aspects.
  - **update** \( s_i \) for all query aspects as \( s_i = s_i + \frac{P[d|q_i]}{\sum_j P[d|q_j]} \).
5. Evaluating Novelty & Diversity

- Traditional effectiveness measures (e.g., MAP and NDCG) and relevance assessments capture neither novelty nor diversity.

- Relevance assessments are collected for (query, document) pairs in isolation, not considering what the user has seen already or to which query aspects the document relates.

- Example: Query *jaguar* with aspects *car* and *cat*

  $$R_1 = \langle d_1, d_1', d_1'', d_2 \rangle \quad R_2 = \langle d_2, d_3, d_3', d_4 \rangle \quad R_3 = \langle d_1, d_3, d_5, d_4 \rangle$$

  Assuming that all documents (e.g., $d_1$) and duplicates (e.g., $d_1'$) are relevant, all three results are considered equally good by existing retrieval effectiveness measures.
5.1. Measuring Diversity

- Agrawal et al. [1], along with IA-SELECT, propose **intent-aware adaptations** of existing retrieval effectiveness measures.

- Let $q_i$ denote the intents (query aspects), $P[q_i|q]$ denote their popularity, and assume that documents have been assessed with regard to their relevance to each intent $q_i$.

- **Example:** **Intent-aware NDCG** (NDCG-IA)

  - Let $\text{NDCG}(q_i, k)$ denote the NDCG at cut-off $k$, assuming $q_i$ as the user’s intent behind the query $q$.

  \[
  \text{NDCG-IA}(q, k) = \sum_i P[q_i|q] \cdot \text{NDCG}(q_i, k)
  \]
Other existing retrieval effectiveness measures (e.g., MAP and MRR) can be made intent-aware using the same approach.

Intent-aware adaptations only capture diversity, i.e., whether different intents are covered by the query result; they do not capture whether what is shown for each of the intents is novel and avoids redundancy.
5.2. Measuring Novelty & Diversity

- Measuring novelty requires **breaking with the assumption** of the PRP that probabilities of relevance are **pairwise independent**

- Clarke et al. [5] propose the **α-nDCG effectiveness measure** which can be instantiated to **capture diversity, novelty, or both**
  - based on the idea of **(information) nuggets** \( n_i \) which can represent any binary property of documents (e.g., query aspect, specific fact)
  - **users** and **documents** represented as **sets of information nuggets**
Probability $P[n_i \in u]$ that nugget $n_i$ is of interest to user $u$
- assumed constant $\gamma$ (e.g., uniform across all nuggets)

Probability $P[n_i \in d]$ that document $d$ is relevant to $n_i$
- obtained from relevance judgment $J(d,i)$ as
  \[
P[n_i \in d] = \begin{cases} 
\alpha & : J(d, i) = 1 \\
0 & : \text{otherwise}
\end{cases}
\]
  with parameter $\alpha$ reflecting trust in reviewers’ assessments

Probability that document $d$ is relevant to user $u$ is
\[
P[R = 1 | u, d] = 1 - \prod_{i=1}^{m} (1 - P[n_i \in u] P[n_i \in d])
\]
α-nDCG

- Probability $P[n_i \in u]$ that nugget $n_i$ is of interest to user $u$
  - assumed constant $\gamma$ (e.g., uniform across all nuggets)
- Probability $P[n_i \in d]$ that document $d$ is relevant to $n_i$
  - obtained from relevance judgment $J(d,i)$ as
    $$P[n_i \in d] = \begin{cases} \alpha & : J(d, i) = 1 \\ 0 & : \text{otherwise} \end{cases}$$
    with parameter $\alpha$ reflecting trust in reviewers’ assessments
- Probability that document $d$ is relevant to user $u$ is
  $$P[R = 1 | u, d] = 1 - \prod_{i=1}^{m} (1 - \gamma \alpha J(d, i))$$
\( \alpha \)-nDCG

- Probability that nugget \( n_i \) is **still of interest to user** \( u \), after having seen documents \( d_1, \ldots, d_{k-1} \)

\[
P[n_i \in u \mid d_1, \ldots, d_{k-1}] = P[n_i \in u] \prod_{j=1}^{k-1} P[n_i \not\in d_j]
\]

- Probability that user sees a **relevant document at rank** \( k \), after having seen documents \( d_1, \ldots d_{k-1} \)

\[
P[R_k = 1 \mid u, d_1, \ldots, d_k] = 1 - \prod_{i=1}^{m} (1 - P[n_i \in u \mid d_1, \ldots, d_{k-1}] P[n_i \in d_k])
\]
α-nDCG

- α-NDCG uses probabilities $P[R_k=1|u,d_1,\ldots,d_k]$ as gain values $G[j]$.

$$DCG[k] = \sum_{j=1}^{k} \frac{G[j]}{\log_2(1+j)}$$

- Finding the ideal gain vector required to compute the idealized DCG for normalization is $NP$-hard (reduction from VERTEX COVER).

- In practice, the idealized DCG, required to obtain nDCG, is approximated by selecting documents using a greedy algorithm.
5.3. TREC Diversity Task

- **Diversity task** within **TREC Web Track** 2009 – 2012
  - **ClueWeb09** as document collection (1 billion web pages)
  - ~50 ambiguous/faceted topics per year

```xml
<topic number="155" type="faceted">
  <query>last supper painting</query>
  <description>
    Find a picture of the Last Supper painting by Leonardo da Vinci.
  </description>
  <subtopic number="1" type="nav">
    Find a picture of the Last Supper painting by Leonardo da Vinci.
  </subtopic>
  <subtopic number="2" type="nav">
    Are tickets available online to view da Vinci’s Last Supper in Milan, Italy?
  </subtopic>
  <subtopic number="3" type="inf">
    What is the significance of da Vinci’s interpretation of the Last Supper in Catholicism?
  </subtopic>
</topic>
```

- Effectiveness measure: \( \alpha \)-nDCG@k and MAP-IA among others
5.3. TREC Diversity Task

- **Diversity task** within TREC Web Track 2009 – 2012

- ClueWeb09 as document collection (1 billion web pages)

- ~50 ambiguous/faceted topics per year

```
<topic number="162" type="ambiguous">
<query>dnr</query>
<description>
  What are "do not resuscitate" orders and how do you get one in place?
</description>
<subtopic number="1" type="inf">
  What are "do not resuscitate" orders and how do you get one in place?
</subtopic>
<subtopic number="2" type="nav">
  What is required to get a hunting license online from the Michigan Department of Natural Resources?
</subtopic>
<subtopic number="3" type="inf">
  What are the Maryland Department of Natural Resources’ regulations for deer hunting?
</subtopic>
</topic>
```

- effectiveness measure: $\alpha$-nDCG@k and MAP-IA among others
Dang and Croft [9] report the following results based on TREC Diversity Track 2009 + 2010, using either the specified subtopics or query suggestions, and comparing

- **Query likelihood** based on unigram language model with Dirichlet smoothing
- **Maximum Marginal Relevance**
- **xQuAD**
- **PM-1 / PM-2**

### Table 2: Performance of all techniques in several standard retrieval tasks

<table>
<thead>
<tr>
<th>Sub-topics</th>
<th>α-NDCG</th>
<th>Prec-IA</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Query-likelihood</strong></td>
<td>0.2979</td>
<td>0.1146</td>
</tr>
<tr>
<td>MMR</td>
<td>0.2963</td>
<td>0.1221</td>
</tr>
<tr>
<td>xQuAD</td>
<td>0.3300&lt;sub&gt;Q,M&lt;/sub&gt;</td>
<td>0.1190</td>
</tr>
<tr>
<td>PM-1</td>
<td>0.3076</td>
<td>0.1140</td>
</tr>
<tr>
<td>PM-2</td>
<td>0.3473&lt;sup&gt;**&lt;/sup&gt;</td>
<td>0.1197</td>
</tr>
<tr>
<td><strong>Suggestions</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Query-likelihood</td>
<td>0.2875</td>
<td>0.1095</td>
</tr>
<tr>
<td>MMR</td>
<td>0.2926</td>
<td>0.1108</td>
</tr>
<tr>
<td>xQuAD</td>
<td>0.2995</td>
<td>0.1089</td>
</tr>
<tr>
<td>PM-1</td>
<td>0.2870</td>
<td>0.0929&lt;sup&gt;†&lt;/sup&gt;</td>
</tr>
<tr>
<td>PM-2</td>
<td>0.3200</td>
<td>0.1123&lt;sup&gt;**&lt;/sup&gt;</td>
</tr>
<tr>
<td><strong>WT-2009 Best (uogTrDYCseB) [10]</strong></td>
<td>0.3081</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>Sub-topics</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Query-likelihood</td>
<td>0.3236</td>
<td>0.1713</td>
</tr>
<tr>
<td>MMR</td>
<td>0.3349&lt;sub&gt;Q&lt;/sub&gt;</td>
<td>0.1740</td>
</tr>
<tr>
<td>xQuAD</td>
<td>0.4074&lt;sub&gt;Q,M&lt;/sub&gt;</td>
<td>0.2028</td>
</tr>
<tr>
<td>PM-1</td>
<td>0.4323&lt;sup&gt;†&lt;/sup&gt;&lt;sub&gt;Q,M&lt;/sub&gt;</td>
<td>0.1827</td>
</tr>
<tr>
<td>PM-2</td>
<td>0.4546&lt;sup&gt;†&lt;/sup&gt;&lt;sub&gt;Q,M&lt;/sub&gt;</td>
<td>0.2030</td>
</tr>
<tr>
<td><strong>Suggestions</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Query-likelihood</td>
<td>0.3268</td>
<td>0.1730</td>
</tr>
<tr>
<td>MMR</td>
<td>0.3361&lt;sub&gt;Q&lt;/sub&gt;</td>
<td>0.1746</td>
</tr>
<tr>
<td>xQuAD</td>
<td>0.3582&lt;sub&gt;Q,M&lt;/sub&gt;</td>
<td>0.1785</td>
</tr>
<tr>
<td>PM-1</td>
<td>0.3664&lt;sup&gt;†&lt;/sup&gt;</td>
<td>0.1654</td>
</tr>
<tr>
<td>PM-2</td>
<td>0.4374&lt;sup&gt;†&lt;/sup&gt;&lt;sub&gt;Q,M&lt;/sub&gt;</td>
<td>0.1841</td>
</tr>
<tr>
<td><strong>WT-2010 Best (uogTrB67xS) [11]</strong></td>
<td>0.4178</td>
<td>N/A</td>
</tr>
</tbody>
</table>
Summary

- **Novelty** reflects how well the returned results avoid **redundancy**.
- **Diversity** reflects how well the returned results resolve **ambiguity**.
- **Probability ranking principle** and its **underlying assumptions** need to be **revised** when aiming for novelty and/or diversity.
- **Implicit methods** for novelty and/or diversity operate directly on the **document contents** without representing query aspects.
- **Explicit methods** for novelty and/or diversity rely on an explicit representation of query aspects (e.g., as query suggestions).
- Standard effectiveness measures do neither capture novelty nor diversity; **intent-aware measures** capture diversity; **cascade measures** (e.g., α-nDCG) can also capture novelty.
# References


References


