Advanced Topics in Information Retrieval

5. Diversity & Novelty

Vinay Setty
(vsetty@mpi-inf.mpg.de)

Jannik Strötgen
(jtroetge@mpi-inf.mpg.de)

max planck institut
informatik
MAX-PLANCK-GESELLSCHAFT

Outline

5.1. Why Novelty & Diversity?

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

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

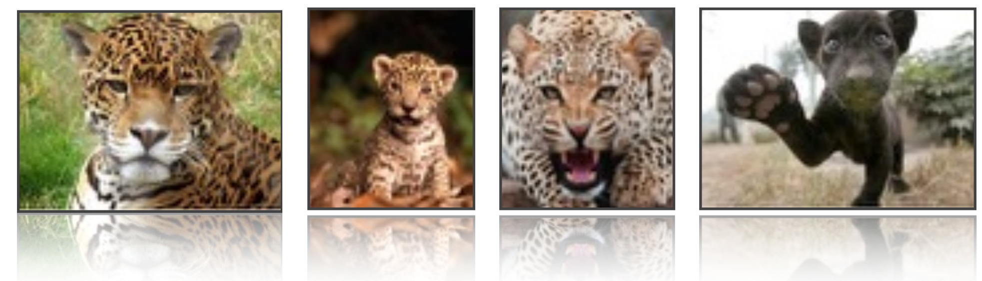
- Bernstein and Zobel [2] identify **near duplicates** in TREC GOV2; mean **MAP dropped by 20.2%** when treating them as **irrelevant** and **increased by 16.0%** when **omitting them** from results

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

5.1. Why Novelty & Diversity?

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

panthera onca ——— ? ———→



- **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 → ? →

- **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 (Beyond Independent Relevance) [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]

Outline

5.1. Why Novelty & Diversity?

5.2. Probability Ranking Principle Revisited

5.3. Implicit Diversification

5.4. Explicit Diversification

5.5. Evaluating Novelty & Diversity

5.2. 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**

max planck institut informatik

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**

Outline

5.1. Why Novelty & Diversity?

5.2. Probability Ranking Principle Revisited

5.3. Implicit Diversification

5.4. Explicit Diversification

5.5. Evaluating Novelty & Diversity

5.3. Implicit Diversification

- **Implicit diversification methods** do not represent query aspects explicitly and instead operate directly on **document contents and their (dis)similarity**

5.3.1. 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**

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

with λ as a **tunable parameter** controlling relevance vs. novelty and *sim* a **similarity measure** (e.g., cosine similarity) between queries and documents

5.3.2. 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

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

  - with value_R as a measure of **relevance** to the query (e.g., the likelihood of generating the query q from θ_i),
  - 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

max planck institut informatik

Beyond Independent Relevance

- 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 θ_O be a language model estimated from **documents** $d_1, \ldots, d_{i-1}$
  - let θ_B be a **background language** model estimated from the **collection**
  - the **log-likelihood** of generating d_i from a mixture of the two is

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

  - the parameter value λ 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

Outline

5.4. 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**

5.4.1. Intent-Aware Selection

- Agrawal et al. [1] model **query aspects as categories** (e.g., from a topic taxonomy such as the Open Directory Project (https://www.dmoz.org))
  - 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]

max planck institut informatik

MAX-PLANCK-GESELLSCHAFT

Intent-Aware Selection

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

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

Intent-Aware Selection

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

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

prob, that d satisfies
query q with category c

max planck institut
informatik

Intent-Aware Selection

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

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

max planck institut informatik

Intent-Aware Selection

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

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

prob. that d fails to satisfy
query q with category c

max planck institut
informatik

Intent-Aware Selection

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

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

Intent-Aware Selection

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

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

prob. that all d in S fail to satisfy query q with category c

Intent-Aware Selection

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

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

max planck institut
informatik

MAX-PLANCK-GESELLSCHAFT

Intent-Aware Selection

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

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

prob, that at least one d in S to satisfies query q
with category c

max planck institut
informatik

Intent-Aware Selection

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

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

max planck institut informatik

MAX-PLANCK-GESELLSCHAFT

Intent-Aware Selection

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

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

prob that a set of
documents S satisfies
the "average" user who
issues query q covering
all categories

max planck institut
informatik

MAX-PLANCK-GESELLSCHAFT

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

IA-SELECT (NP-Hard Problem)

- Theorem: Finding the set S of size k that maximizes

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

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**

$$\sum_c \mathrm{P}[\neg c \mid S] \cdot \mathrm{P}[q \mid d] \cdot \mathrm{P}[c \mid d]$$

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

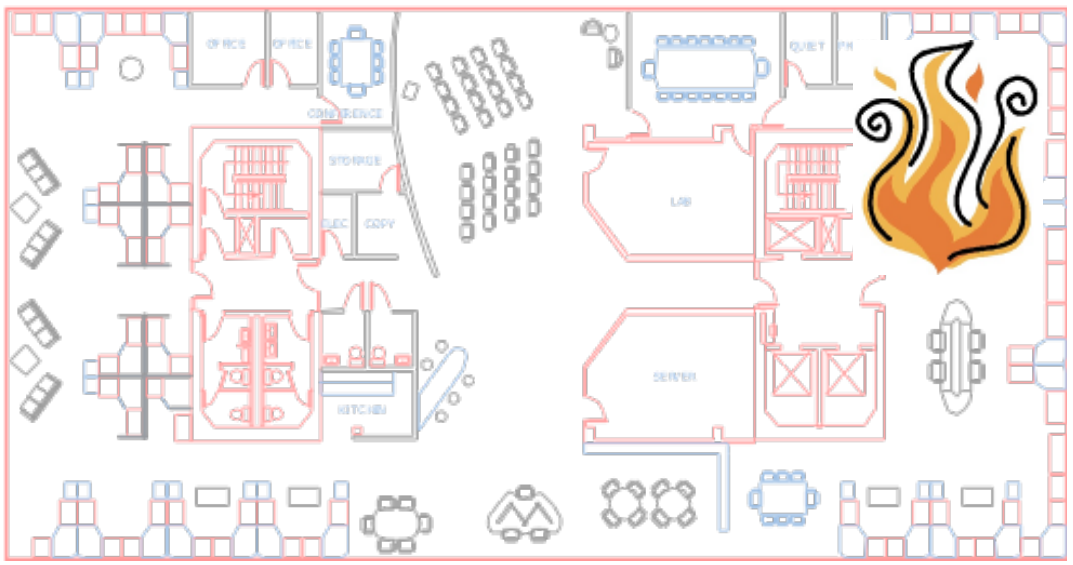
$$\mathrm{P}[\neg c \mid S] = \prod_{d \in S} (1 - \mathrm{P}[q \mid d] \cdot \mathrm{P}[c \mid d])$$

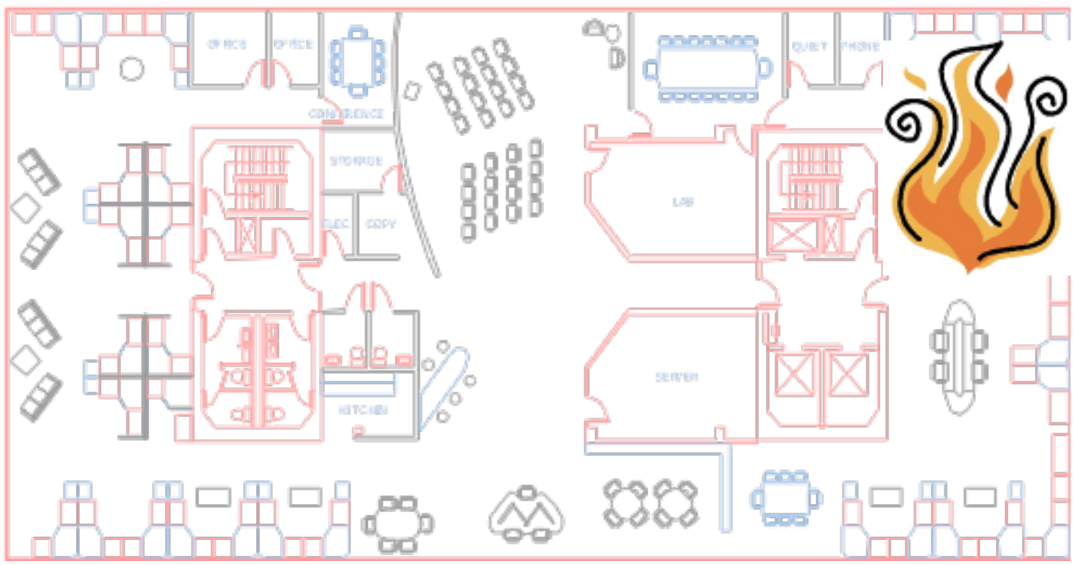
which is initialized as P[c|q]

Submodularity & Approximation

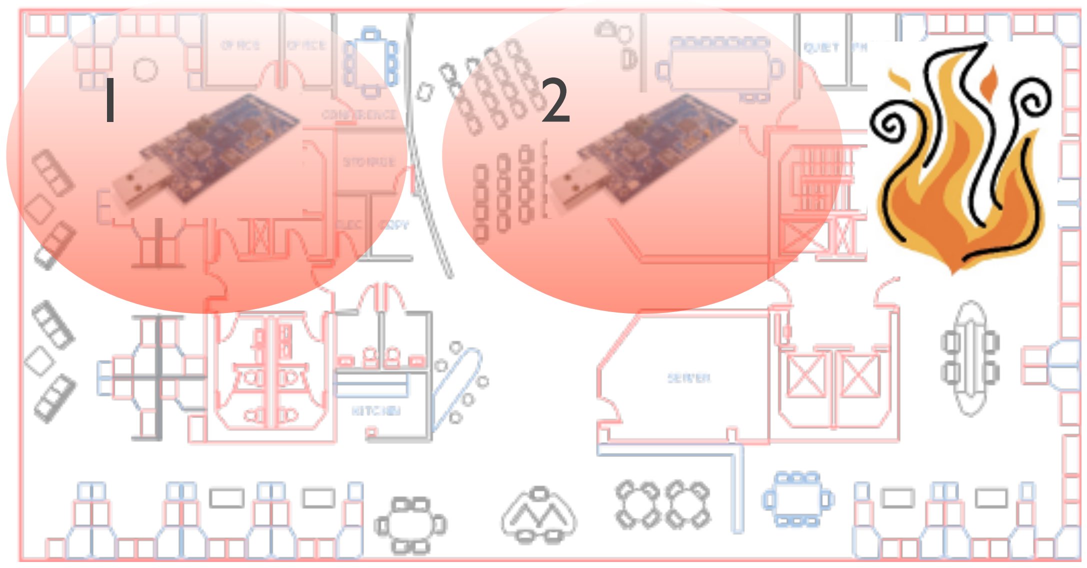
- Definition: Given a finite ground set N, a function $f: 2^N \longrightarrow R$ is **submodular** if and only if for all sets $S, T \subseteq N$
  - such that $S \subseteq T$,
  - and $d \in N \setminus T$, $f(S \cup \{d\}) - f(S) \geq f(T \cup \{d\}) - f(T)$

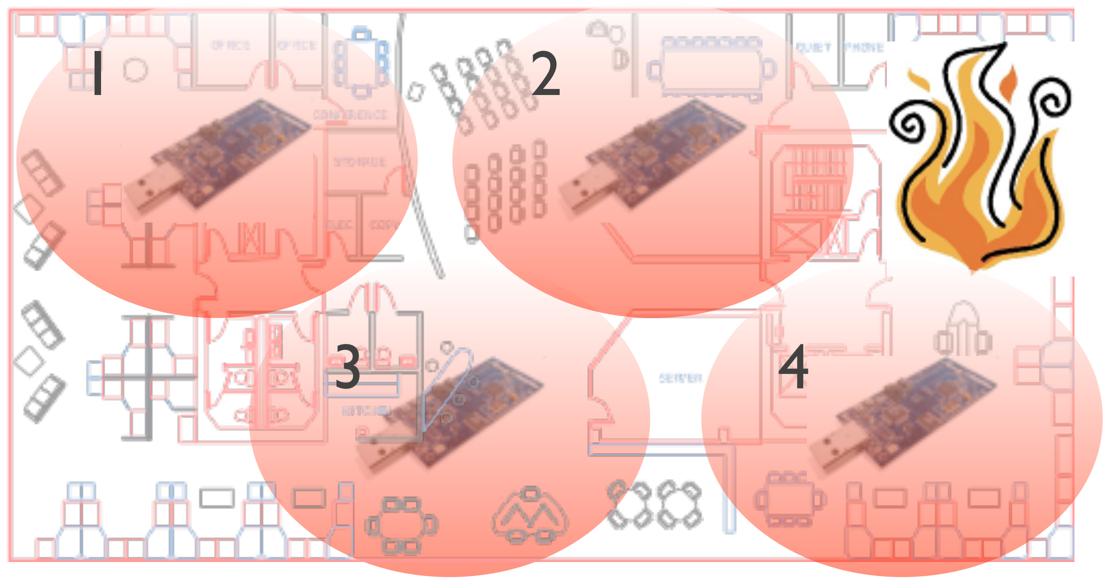
Submodularity Gain Example



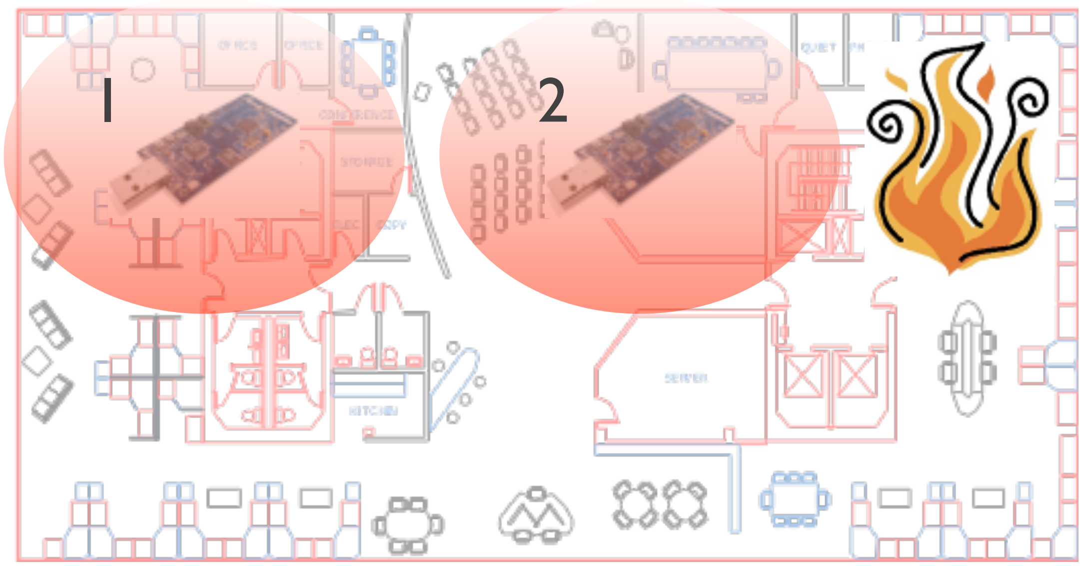


Submodularity Gain Example

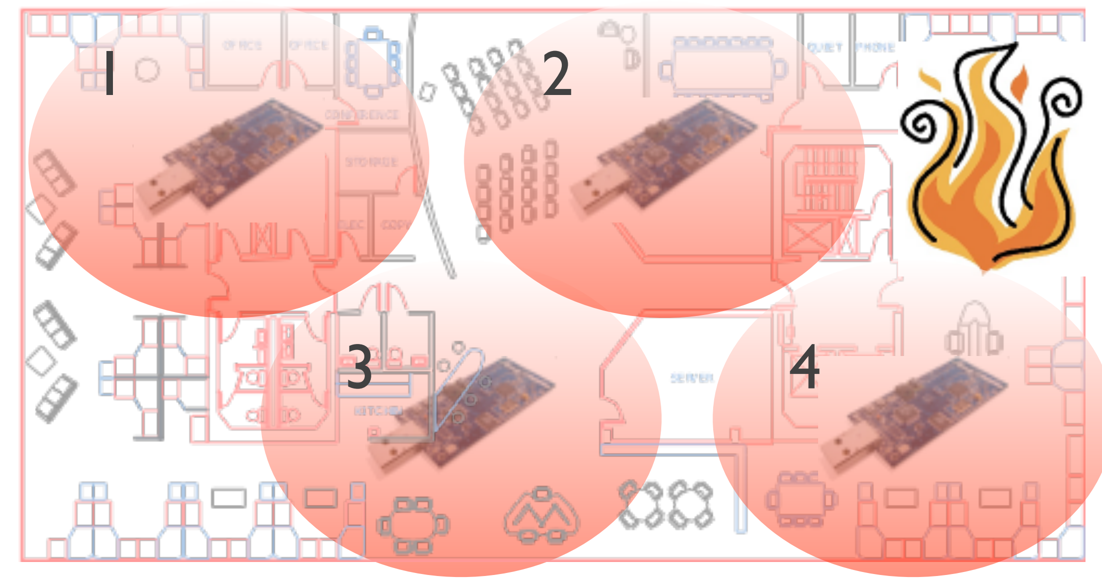




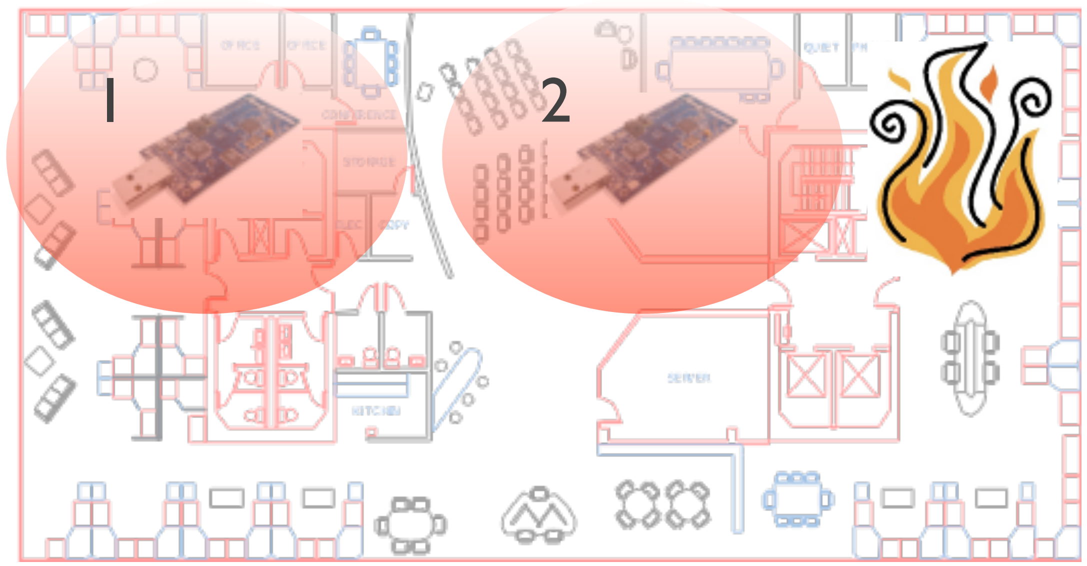
Submodularity Gain Example



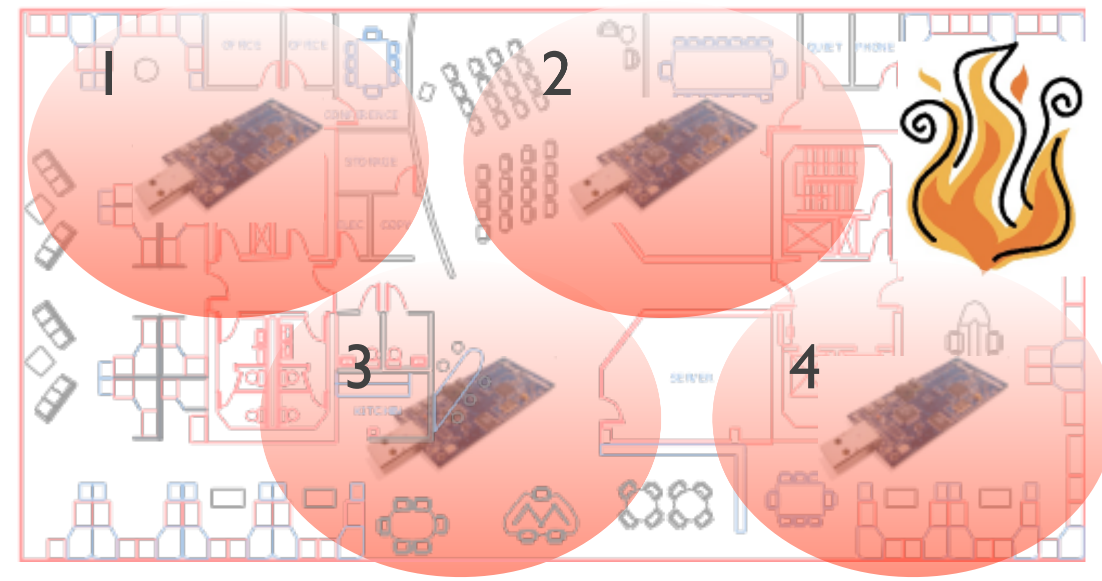
A = {1,2}



Submodularity Gain Example

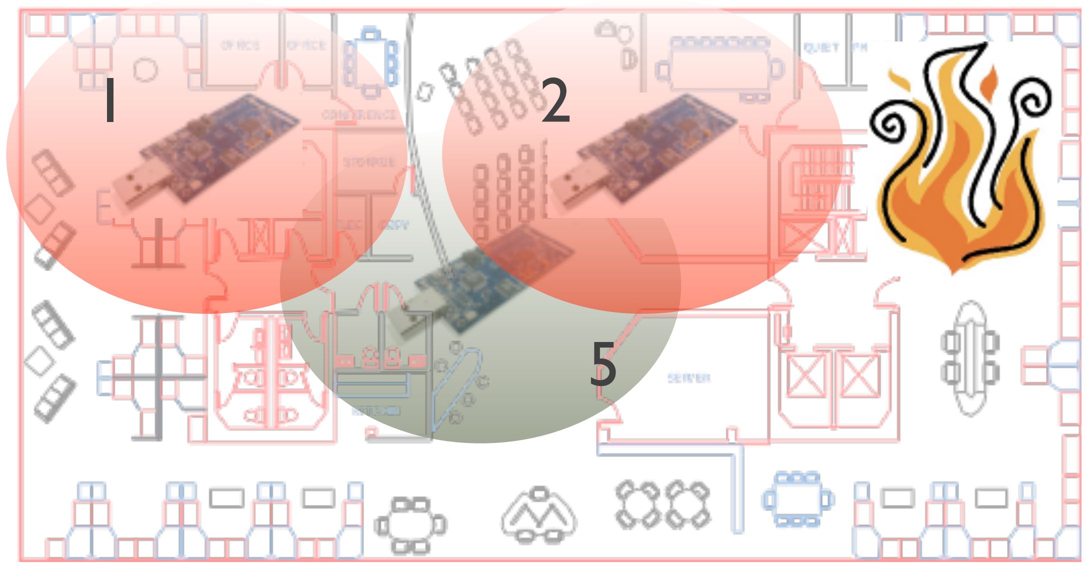


A = {1,2}

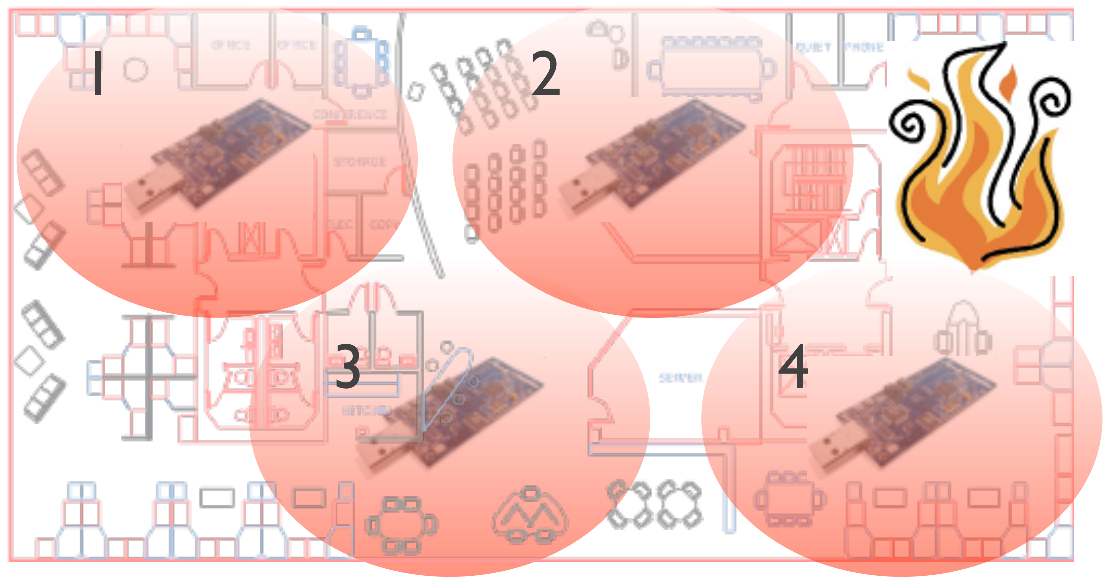


B = {1,2,3,4}

Submodularity Gain Example

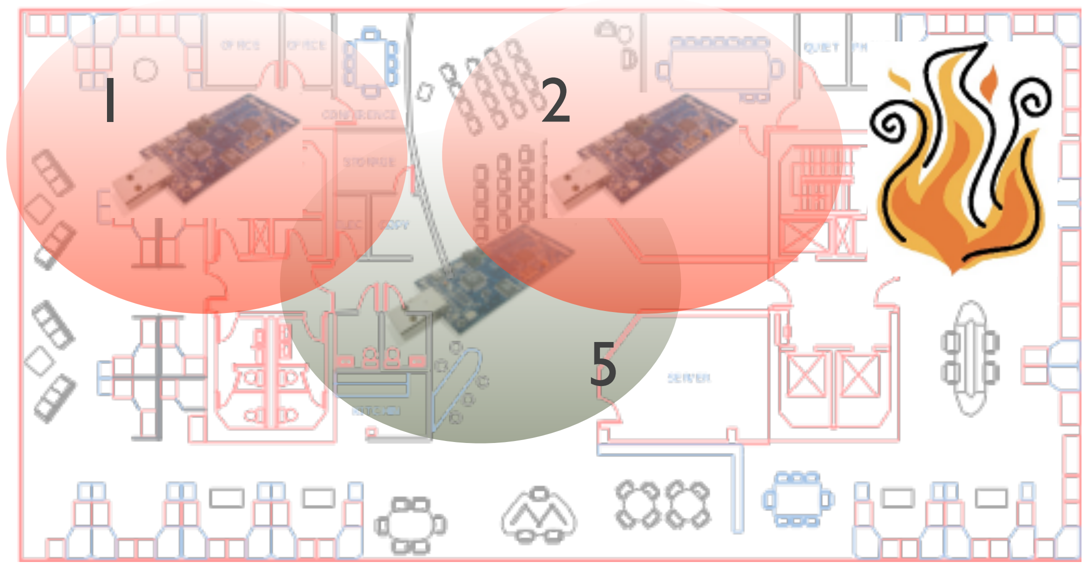


A = {1,2}

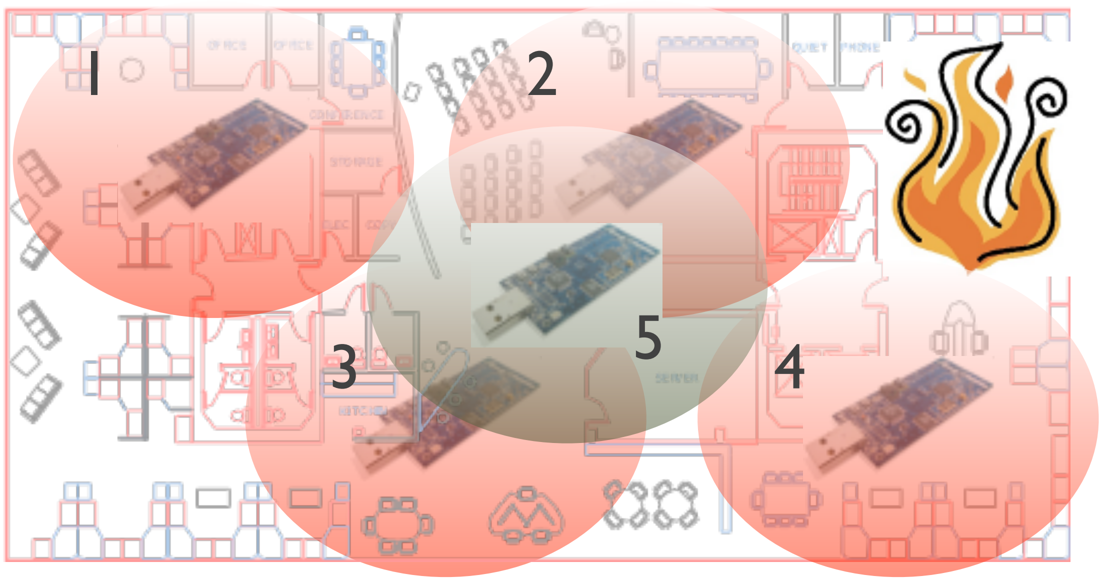


B = {1,2,3,4}

Submodularity Gain Example

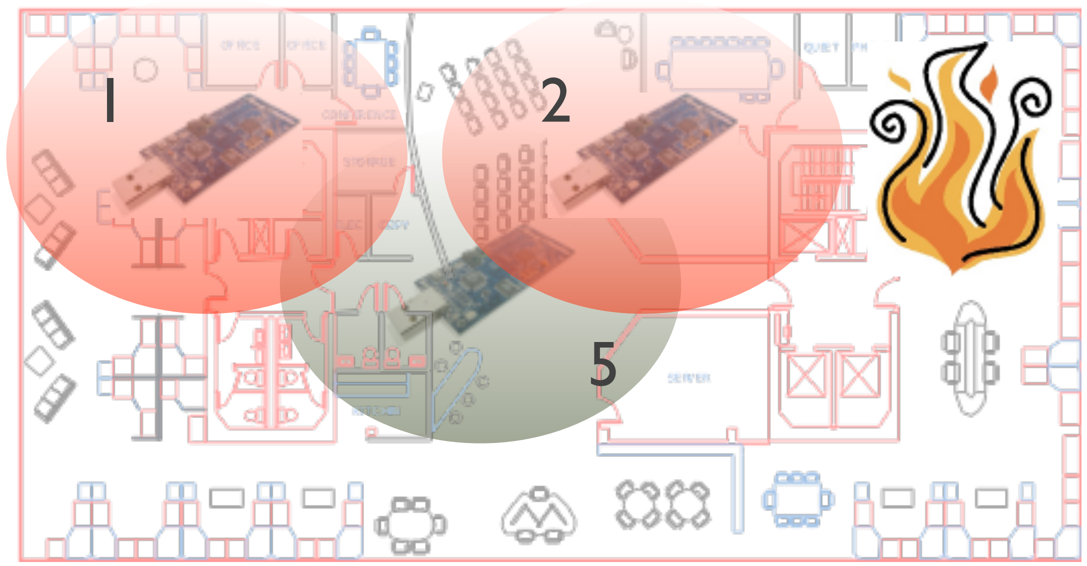


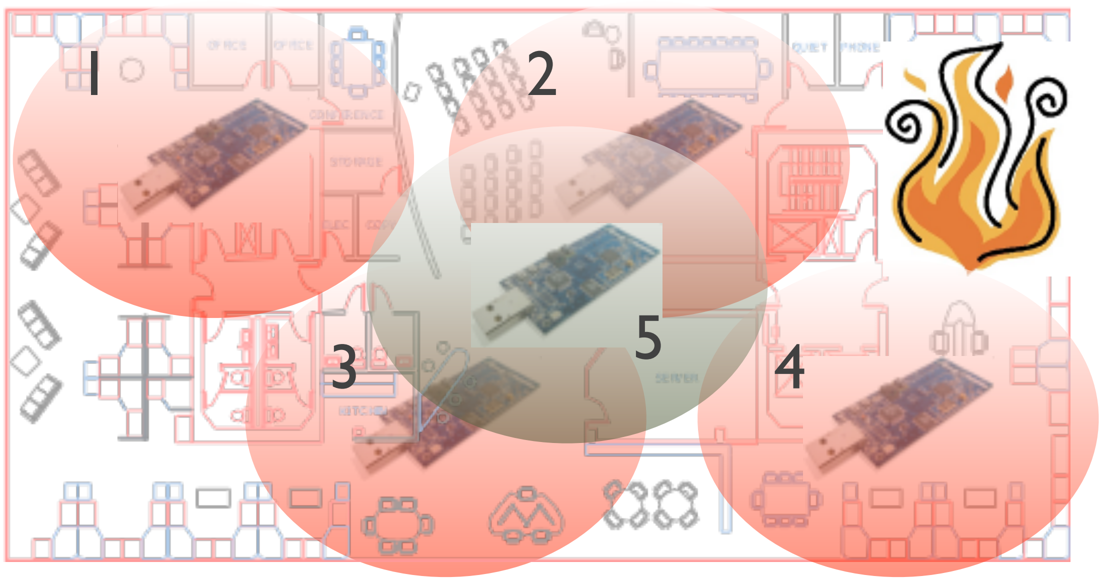
A = {1,2}



B = {1,2,3,4}

Submodularity Gain Example





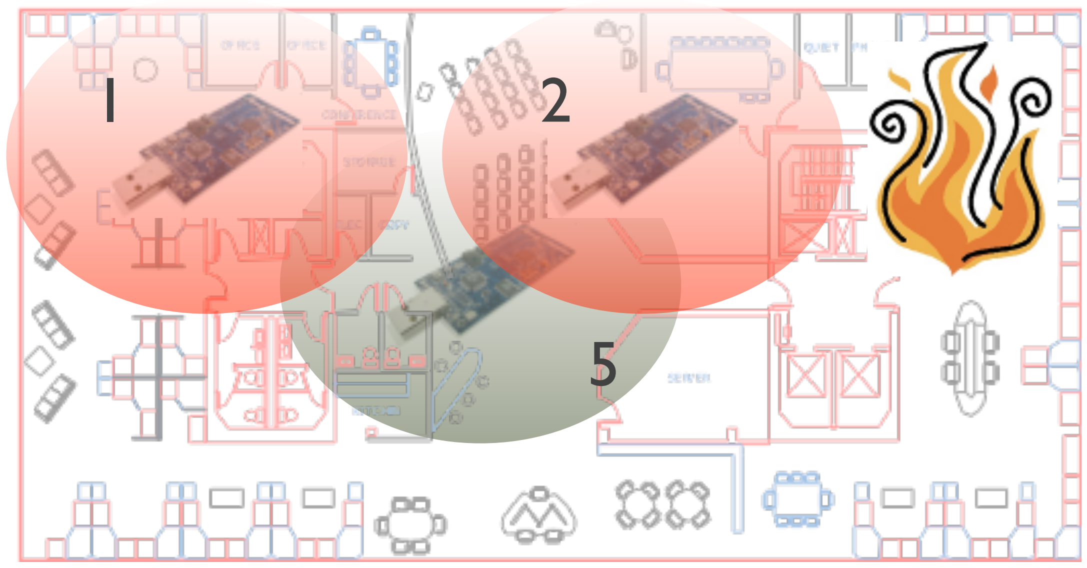
A = {1,2}

B = {1,2,3,4}

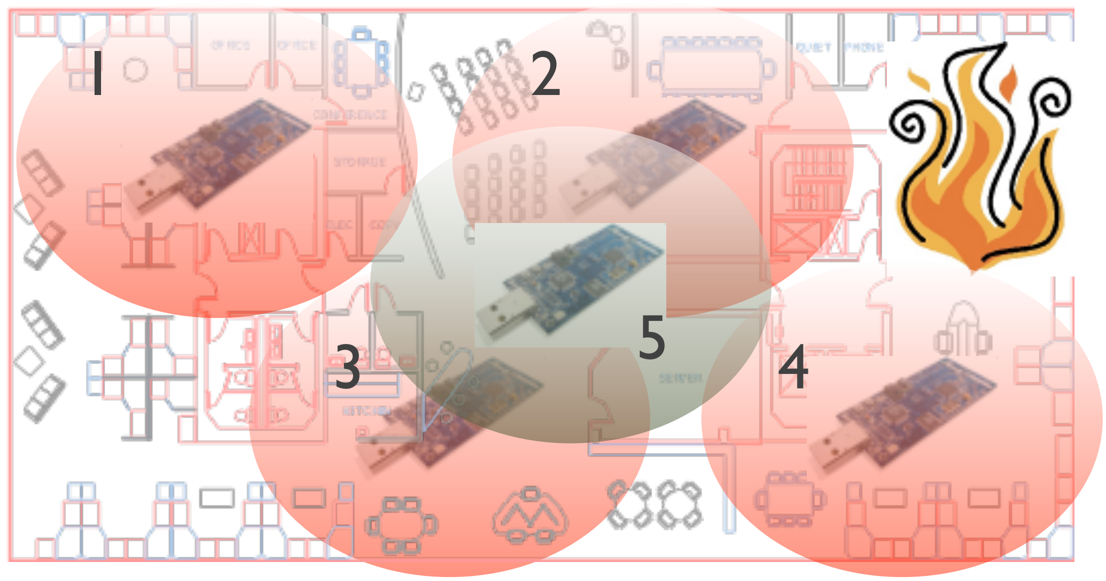
$A \subseteq B$

$f(A \cup 5) \geq f(B \cup 5)$

Submodularity Gain Example



A = {1,2}



B = {1,2,3,4}

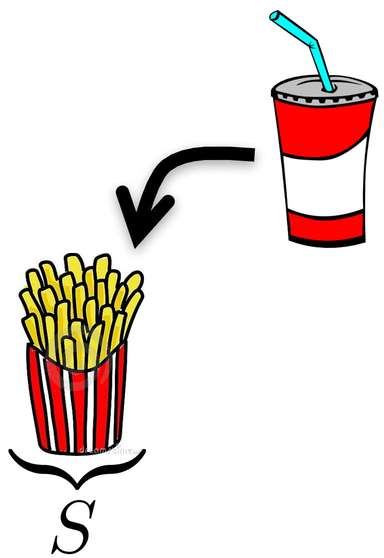
$A \subseteq B$

$f(A \cup 5) \geq f(B \cup 5)$

Diminishing marginal gains

max planck institut
informatik

Submodularity Cost Example

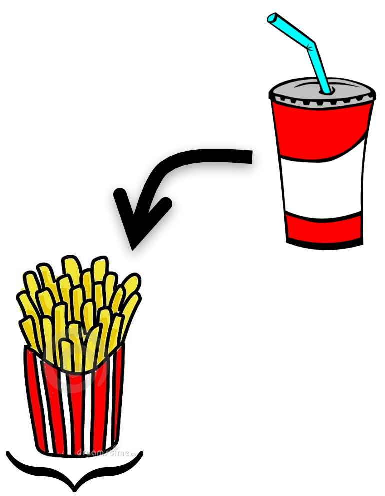


$$F(S \cup s) - F(S)$$

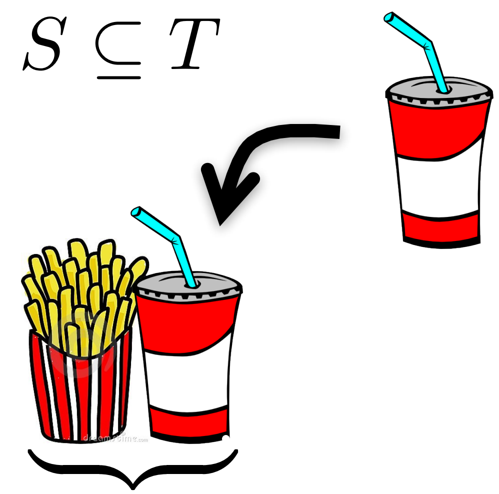
extra cost:
one drink

Submodularity Cost Example

$S \subseteq T$



S



T

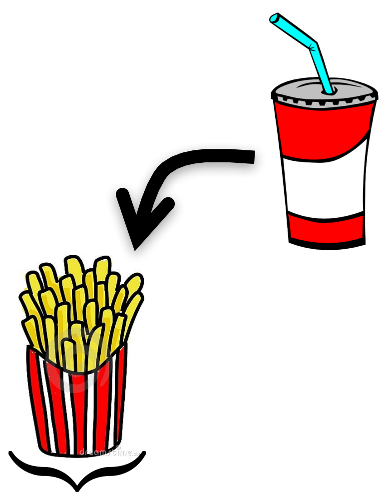
$$F(S \cup s) - F(S) \geq F(T \cup s) - F(T)$$

extra cost:
one drink

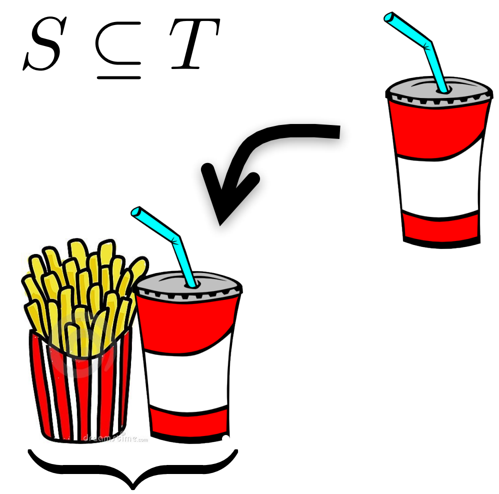
extra cost:
free refill ☺

Submodularity Cost Example

$S \subseteq T$



S



T

$$F(S \cup s) - F(S) \geq F(T \cup s) - F(T)$$

extra cost:
one drink

extra cost:
free refill ☺

Diminishing marginal costs

max planck institut
informatik

MAX-PLANCK-GESELLSCHAFT

IA-Select is Submodular

- 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') ≥ (1 - 1/e) f(S*)
- Corollary: IA-SELECT is (1-1/e)-approximation algorithm follows from proof in [12]

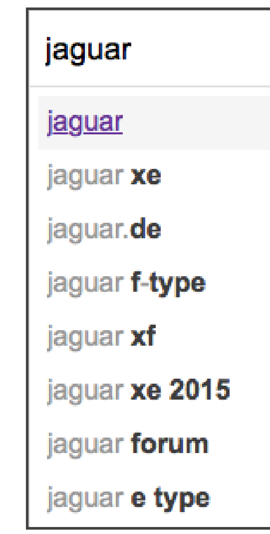
max planck institut informatik

MAX-PLANCK-GESELLSCHAFT

5.4.2. 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



$$(1-\lambda)\ \mathrm{P}[\,d \mid q\,] + \lambda\, P[\,d, \neg S \mid q\,]$$

where P[d|q] is the document likelihood and captures **relevance** and P[d,¬S|q] is the probability that d covers a query aspect not yet covered by documents in S and captures **diversity**

xQuAD

- **Probability P[d,¬S|q]** can be **decomposed into**

$$\sum_i \mathrm{P}\left[\, \neg S \mid q_i \,\right] \, \mathrm{P}\left[\, q_i \mid q \,\right]$$

- **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

$$\mathrm{P}\left[\, \neg S \mid q_i \,\right] = \prod_{d \in S} \left(1 - \mathrm{P}\left[\, d \mid q_i \,\right]\right)$$

max planck institut informatik

MAX-PLANCK-GESELLSCHAFT

IA-SELECT and XQUAD Limitations

max planck institut
informatik

IA-SELECT and xQUAD Limitations

- Redundancy-based methods (IA-SELECT and xQUAD) degenerate

IA-SELECT and XQUAD Limitations

- 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 and XQUAD Limitations

- 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

IA-SELECT and XQUAD Limitations

- 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

MAX-PLANCK-GESELLSCHAFT mpii max planck institut informatik

5.4.3. 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

- <u>Example</u>: Query jaguar refers to query aspect car with 75% probability and to query aspect cat with 25% probability

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

  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

$$\frac{v_i}{2\,s_i + 1}$$

    - increment number of seats s_i allocated to party p_i

Sainte-Laguë Method Example

s_i $\cdot$ $$\frac{v_i}{2\,s_i + 1}$$

Sainte-Laguë Method Example

$$\frac{v_i}{2\,s_i + 1}$$

s_i	
0	49
0	22
0	15
0	1

Num votes received

Sainte-Laguë Method Example

$$\frac{v_i}{2\,s_i + 1}$$

s_i	
1	<u>*49*</u>
0	22
0	15
0	1

Num votes received

max planck institut informatik

Sainte-Laguë Method Example

s_i		$\frac{v_i}{2\,s_i+1}$
1	49	49/3
0	22	22
0	15	15
0	1	1

Num votes received

Sainte-Laguë Method Example

s_i		$\frac{v_i}{2\,s_i+1}$
1	49	49/3
1	22	22
0	15	15
0	1	1

Num votes received

max planck institut informatik

Sainte-Laguë Method Example

$$\frac{v_i}{2\,s_i + 1}$$

s_i			
1	<u>*49*</u>	49/3	<u>*16.33*</u>
1	22	<u>22</u>	22/3
0	15	15	15
0	1	1	1

Num votes received

Sainte-Laguë Method Example

s_i	$\frac{v_i}{2\,s_i + 1}$		
2	*49*	49/3	*16.33*
1	22	22	22/3
0	15	15	15
0	1	1	1

Num votes received

Sainte-Laguë Method Example

s_i	$\frac{v_i}{2\,s_i+1}$			
2	*49*	49/3	*16.33*	49/5
1	22	*22*	22/3	7.33
0	15	15	15	*15*
0	1	1	1	1

Num votes received

Sainte-Laguë Method Example

$$\frac{v_i}{2\,s_i + 1}$$

s_i				
2	49	49/3	16.33	49/5
1	22	22	22/3	7.33
1	15	15	15	15
0	1	1	1	1

Num votes received

max planck institut informatik

MAX-PLANCK-GESELLSCHAFT

Sainte-Laguë Method Example

s_i	$\frac{v_i}{2\,s_i+1}$				
2	*49*	49/3	*16.33*	49/5	*9.8*
1	22	22	22/3	7.33	7.33
1	15	15	15	*15*	5
0	1	1	1	1	1

Num votes received

Sainte-Laguë Method Example

s_i	$\frac{v_i}{2\,s_i+1}$				
3	*49*	49/3	*16.33*	49/5	*9.8*
1	22	*22*	22/3	7.33	7.33
1	15	15	15	*15*	5
0	1	1	1	1	1

Num votes received

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

$$\frac{v_i}{2\, s_i + 1}$$

    - **select document** d having the highest score

$$\lambda \cdot \frac{v_i}{2\, s_i + 1} \cdot \mathrm{P}\left[\, d \mid q_i \,\right] + (1-\lambda) \cdot \sum_{j \neq i} \frac{v_j}{2\, s_j + 1} \cdot \mathrm{P}\left[\, d \mid q_j \,\right]$$

      with parameter λ trading off relatedness to aspect q_i vs. all other aspects

    - update s_i for all query aspects as $s_i = s_i + \dfrac{\mathrm{P}\left[\, d \mid q_i \,\right]}{\sum_j \mathrm{P}\left[\, d \mid q_j \,\right]}$

Outline

5.5. 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

Recap of nDCG

Recap of nDCG

- rel_i is the relevance grade given to a document at position rank i usually ranges from 0 to 3 or 5

max planck institut
informatik

Recap of nDCG

- rel_i is the relevance grade given to a document at position rank i usually ranges from 0 to 3 or 5
- What is CG (Cumulative gain)?

Recap of nDCG

- *rel*$_i$ is the relevance grade given to a document at position rank i usually ranges from 0 to 3 or 5
- What is CG (Cumulative gain)?

Cumulative sum of relevance scores up to a rank position p

$$\mathrm{CG_p} = \sum_{i=1}^{p} rel_i$$

Recap of nDCG

- *rel*$_i$ is the relevance grade given to a document at position rank i usually ranges from 0 to 3 or 5
- What is CG (Cumulative gain)?
  - What is the problem with it?

Cumulative sum of relevance scores up to a rank position p

$$\mathrm{CG_p} = \sum_{i=1}^{p} rel_i$$

max planck institut informatik

MAX-PLANCK-GESELLSCHAFT

Recap of nDCG

- rel_i is the relevance grade given to a document at position rank i usually ranges from 0 to 3 or 5
- What is CG (Cumulative gain)?
  - What is the problem with it?
- What is DCG (Discounted CG)?

Cumulative sum of relevance scores up to a rank position p

$$\mathrm{CG_p} = \sum_{i=1}^{p} rel_i$$

max planck institut informatik

Recap of nDCG

- rel_i is the relevance grade given to a document at position rank i usually ranges from 0 to 3 or 5
- What is CG (Cumulative gain)?
  - What is the problem with it?
- What is DCG (Discounted CG)?

Cumulative sum of relevance scores up to a rank position p

$$\mathrm{CG_p} = \sum_{i=1}^{p} rel_i$$

CG_p + penalizes highly relevant documents occurring at lower ranks

$$\mathrm{DCG_p} = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2(i)}$$

Recap of nDCG

- rel_i is the relevance grade given to a document at position rank i usually ranges from 0 to 3 or 5
- What is CG (Cumulative gain)?
  - What is the problem with it?
- What is DCG (Discounted CG)?
  - What is the problem with it?

Cumulative sum of relevance scores up to a rank position p

$$\mathrm{CG_p} = \sum_{i=1}^{p} rel_i$$

CG_p + penalizes highly relevant documents occurring at lower ranks

$$\mathrm{DCG_p} = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2(i)}$$

max planck institut informatik

Recap of nDCG

- rel_i is the relevance grade given to a document at position rank i usually ranges from 0 to 3 or 5
- What is CG (Cumulative gain)?
  - What is the problem with it?

Cumulative sum of relevance scores up to a rank position p

$$\mathrm{CG_p} = \sum_{i=1}^{p} rel_i$$

- What is DCG (Discounted CG)?
  - What is the problem with it?

CG_p + penalizes highly relevant documents occurring at lower ranks

$$\mathrm{DCG_p} = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2(i)}$$

- What is nDCG (Normalized DCG)?

Recap of nDCG

- *rel*$_i$ is the relevance grade given to a document at position rank i usually ranges from 0 to 3 or 5
- What is CG (Cumulative gain)?
  - What is the problem with it?
- What is DCG (Discounted CG)?
  - What is the problem with it?
- What is nDCG (Normalized DCG)?

Cumulative sum of relevance scores up to a rank position p

$$\mathrm{CG_p} = \sum_{i=1}^{p} rel_i$$

CG$_p$ + penalizes highly relevant documents occurring at lower ranks

$$\mathrm{DCG_p} = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2(i)}$$

IDCGp (Ideal *DCG$_p$*) is the ordering of the documents that maximizes *DCG$_p$*

$$\mathrm{nDCG_p} = \frac{DCG_p}{IDCG_p}$$

Recap of nDCG

- rel_i is the relevance grade given to a document at position rank i usually ranges from 0 to 3 or 5
- What is CG (Cumulative gain)?
  - What is the problem with it?
- What is DCG (Discounted CG)?
  - What is the problem with it?
- What is nDCG (Normalized DCG)?
  - What is the problem with it?

Cumulative sum of relevance scores up to a rank position p

$$\mathrm{CG_p} = \sum_{i=1}^{p} rel_i$$

CG_p + penalizes highly relevant documents occurring at lower ranks

$$\mathrm{DCG_p} = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2(i)}$$

$IDCG_p$ (Ideal DCG_p) is the ordering of the documents that maximizes DCG_p

$$\mathrm{nDCG_p} = \frac{DCG_p}{IDCG_p}$$

max planck institut informatik

nDCG Example

- Say we have d1, d2, d3, d4, d5, d6 with $rel_i = 3,2,3,0,1,2$
- Then what is CG_6?

max planck institut informatik

MAX-PLANCK-GESELLSCHAFT

nDCG Example

- Say we have d1, d2, d3, d4, d5, d6 with $rel_i = 3,2,3,0,1,2$
- Then what is CG_6?

$$\mathrm{CG}_6 = \sum_{i=1}^{6} rel_i = 3 + 2 + 3 + 0 + 1 + 2 = 11$$

nDCG Example

- Say we have d1, d2, d3, d4, d5, d6 with $rel_i = 3,2,3,0,1,2$
- Then what is CG_6?

$$\mathrm{CG}_6 = \sum_{i=1}^{6} rel_i = 3 + 2 + 3 + 0 + 1 + 2 = 11$$

$$\mathrm{DCG}_6 = rel_1 + \sum_{i=2}^{6} \frac{rel_i}{\log_2 i} = 3 + (2 + 1.892 + 0 + 0.431 + 0.774) = 8.10$$

nDCG Example

- Say we have d1, d2, d3, d4, d5, d6 with *rel_i = 3,2,3,0,1,2*
- Then what is CG_6?

$$\mathrm{CG}_6 = \sum_{i=1}^{6} rel_i = 3 + 2 + 3 + 0 + 1 + 2 = 11$$

$$\mathrm{DCG}_6 = rel_1 + \sum_{i=2}^{6} \frac{rel_i}{\log_2 i} = 3 + (2 + 1.892 + 0 + 0.431 + 0.774) = 8.10$$

  - IDCG is computed using ideal ordering of relevance grades = 3,3,2,2,1,0
    - $IDCG_6 = 8.69$

nDCG Example

- Say we have d1, d2, d3, d4, d5, d6 with *rel_i = 3,2,3,0,1,2*
- Then what is CG_6?

$$\mathrm{CG}_6 = \sum_{i=1}^{6} rel_i = 3 + 2 + 3 + 0 + 1 + 2 = 11$$

$$\mathrm{DCG}_6 = rel_1 + \sum_{i=2}^{6} \frac{rel_i}{\log_2 i} = 3 + (2 + 1.892 + 0 + 0.431 + 0.774) = 8.10$$

  - IDCG is computed using ideal ordering of relevance grades = 3,3,2,2,1,0
    - $IDCG_6$ = 8.69
  - Then

max planck institut informatik

MAX-PLANCK-GESELLSCHAFT

nDCG Example

- Say we have d1, d2, d3, d4, d5, d6 with *rel_i = 3,2,3,0,1,2*
- Then what is CG_6?

$$\mathrm{CG}_6 = \sum_{i=1}^{6} rel_i = 3+2+3+0+1+2 = 11$$

$$\mathrm{DCG}_6 = rel_1 + \sum_{i=2}^{6} \frac{rel_i}{\log_2 i} = 3+(2+1.892+0+0.431+0.774) = 8.10$$

  - IDCG is computed using ideal ordering of relevance grades = 3,3,2,2,1,0
    - $\mathrm{IDCG}_6 = 8.69$
  - Then

$$\mathrm{nDCG}_6 = \frac{DCG_6}{IDCG_6} = \frac{8.10}{8.69} = 0.932$$

max planck institut informatik

MAX-PLANCK-GESELLSCHAFT

5.5.1. 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 NDCG(q_i, k) denote the NDCG at cut-off k, assuming q_i as the user's intent behind the query q

$$\text{NDCG-IA}(q,k) = \sum_i \mathrm{P}\left[\, q_i \mid q \,\right] \, \mathrm{NDCG}(q_i, k)$$

Intent-Aware Effectiveness Measures

- 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**

5.5.2. 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**

α-nDCG

- Probability $P[n_i \in u]$ that nugget n_i is of interest to user u
  - assumed **constant** γ (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

$$\mathrm{P}\left[\, n_i \in d \,\right] = \begin{cases} \alpha & : \quad J(d,i) = 1 \\ 0 & : \quad \text{otherwise} \end{cases}$$

with parameter α reflecting trust in reviewers' assessments

- Probability that document d is relevant to user u is

$$\mathrm{P}\left[\, R=1 \,|\, u,d \,\right] = 1 - \prod_{i=1}^{m} \left(1 - \gamma \alpha J(d,i)\right)$$

$$\mathrm{P}\left[\, R=1 \,|\, u,d \,\right] = 1 - \prod_{i=1}^{m} \left(1 - \mathrm{P}\left[\, n_i \in u \,\right] \mathrm{P}\left[\, n_i \in d \,\right]\right)$$

α-nDCG for Evaluating Novelty

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

$$P[\, n_i \in u \mid d_1, \ldots, d_{k-1} \,] = P[\, n_i \in u \,] \prod_{j=1}^{k-1} P[\, n_i \notin d_j \,]$$

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

$$P[\, R_k = 1 \mid u, d_1, \ldots, d_k \,] =$$

$$1 - \prod_{i=1}^{m} (1 - P[\, n_i \in u \,] \prod_{j=1}^{k-1} P[\, n_i \notin d_j \,]\, P[\, n_i \in d_k \,])$$

$$= 1 - \prod_{i=1}^{m} (1 - \quad \gamma \quad (1-\alpha)^{r_{i,k-1}} \quad \alpha J(d_k, i))$$

Where, $r_{i,k-1} = \sum_{j=1}^{k-1} J(d_j, i),$

max planck institut informatik

α-nDCG

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

$$G[k] \;=\; \sum_{i=1}^{m} J(d_k, i)(1-\alpha)^{r_{i,k-1}}. \qquad \mathrm{DCG}[k] = \sum_{j=1}^{k} \frac{G[j]}{\log_2(1+j)}$$

- 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**

max planck institut informatik

α-nDCG Example

85: Norwegian Cruise Lines (NCL)

85.1: Name the ships of the NCL.
85.2: What cruise line attempted to take over NCL in 1999?
85.3: What is the name of the NCL's own private island?
85.4: How does NCL rank in size with other cruise lines?
85.5: Why did the Grand Cayman turn away a NCL ship?
85.6: Name so-called theme cruises promoted by NCL.

Ideal ordering:
a-e-g-b-f-c-h-i-j

Document Title	85.1	85.2	85.3	85.4	85.5	85.6	Total
a. Carnival Re-Enters Norway Bidding		X		X			2
b. NORWEGIAN CRUISE LINE SAYS OUTLOOK IS GOOD		X					1
c. Carnival, Star Increase NCL Stake		X					1
d. Carnival, Star Solidify Control							0
e. HOUSTON CRUISE INDUSTRY GETS BOOST WITH...	X					X	2
f. TRAVELERS WIN IN CRUISE TUG-OF-WAR	X						1
g. ARMCHAIR QUARTERBACKS NEED... THIS CRUISE			X				1
h. EUROPE, CHRISTMAS ON SALE	X						1
i. TRAVEL DEALS AND DISCOUNTS							0
j. HAVE IT YOUR WAY ON THIS SHIP							0

Assuming α = 0.5

$$G[k] = \sum_{i=1}^{m} J(d_k, i)(1-\alpha)^{r_{i,k-1}}.$$

$$G = \langle 2, \tfrac{1}{2}, \tfrac{1}{4}, 0, 2, \tfrac{1}{2}, 1, \tfrac{1}{4}, \ldots \rangle.$$

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

$$CG = \langle 2, 2\tfrac{1}{2}, 2\tfrac{3}{4}, 2\tfrac{3}{4}, 4\tfrac{3}{4}, 5\tfrac{1}{4}, 6\tfrac{1}{4}, 6\tfrac{1}{2}, \ldots \rangle.$$

$$DCG = \langle 2, 2.315, 2.440, \ldots \rangle.$$

max planck institut
informatik

5.5.3. 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="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 measures: **α-nDCG@k** and **MAP-IA** among others

5.5.3. 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 Michiga
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 measures: **α-nDCG@k** and **MAP-IA** among others

TREC Diversity Task Results

- 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

		α-NDCG	Prec-IA
Sub-topics	Query-likelihood	0.2979	0.1146
	MMR	0.2963	**0.1221**
	xQuAD	$0.3300_{Q,M}$	0.1190
	PM-1	0.3076	0.1140
	PM-2	$\mathbf{0.3473}^{P}$	0.1197
Suggestions	Query-likelihood	0.2875	0.1095
	MMR	0.2926	0.1108
	xQuAD	0.2995	0.1089
	PM-1	0.2870	0.0929^{X}
	PM-2	**0.3200**	$\mathbf{0.1123}^{P}$
WT-2009 Best (uogTrDYCcsB) [10]		0.3081	N/A
Sub-topics	Query-likelihood	0.3236	0.1713
	MMR	0.3349_{Q}	0.1740
	xQuAD	$0.4074_{Q,M}$	0.2028
	PM-1	$0.4323^{X}_{Q,M}$	0.1827
	PM-2	$\mathbf{0.4546}^{X,P}_{Q,M}$	**0.2030**
Suggestions	Query-likelihood	0.3268	0.1730
	MMR	0.3361_{Q}	0.1746
	xQuAD	$0.3582_{Q,M}$	0.1785
	PM-1	0.3664^{X}	0.1654
	PM-2	$\mathbf{0.4374}^{X,P}_{Q,M}$	**0.1841**
WT-2010 Best (uogTrB67xS) [11]		0.4178	N/A

max planck institut
informatik

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

max planck institut informatik

References

[1] R. Agrawal, S. Gollapudi, A. Halverson, S. Ieong: *Diversifying Search Results*, WSDM 2009

[2] Y. Bernstein and J. Zobel: *Redundant Documents and Search Effectiveness*, CIKM 2005

[3] J. Carbonell and J. Goldstein: *The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing* Summaries, SIGIR 1998

[4] O. Chapelle, D. Metzler, Y. Zhang, P. Grinspan: *Expected Reciprocal Rank for Graded Relevance*, CIKM 2009

[5] C. L. A. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, I. MacKinnon: *Novelty and Diversity in Information Retrieval Evaluation*, SIGIR 2008

[6] C. L. A. Clarke, N. Craswell, I. Soboroff, A. Ashkan: *A Comparative Analysis of Cascade Measures for Novelty and Diversity*, WSDM 2011

References

[7] Van Dang and W. Bruce Croft: *Diversity by Proportionality: An Election-based Approach to Search Result Diversification*, SIGIR 2012

[8] Van Dang and W. Bruce Croft: *Term Level Search Result Diversification*, SIGIR 2013

[9] S. Robertson: *The Probability Ranking Principle in Information Retrieval*, Journal of Documentation 33(4), 1977

[10] R. L. T. Santos, C. Macdonald, I. Ounis: *Exploiting Query Reformulations for Web Search Result Diversification*, WWW 2010

[11] C. Zhai, W. W. Cohen, J. Lafferty: *Beyond Independent Relevance: Methods and Evaluation Metrics for Subtopic Retrieval*, SIGIR 2003

[12] M. Fisher, G. Nemhauser, and L. Wolsey, *"An analysis of approximations for maximizing submodular set functions-I,"* in Polyhedral Combinatorics. Springer Berlin Heidelberg, 1978.