UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI INGEGNERIA
DELL'INFORMAZIONE

Corso di Laurea Magistrale in Ingegneria Informatica

Tesi di Laurea Magistrale

Uno strumento visuale per l'esplorazione dei dati della valutazione dei sistemi di reperimento dell'informazione

Laureando:
Giacomo ROCCO

Relatore:
Dr. Ing. Gianmaria SILVELLO

Anno Accademico 2016/2017

Sommario

La valutazione sperimentale permette di misurare l'efficacia dei sistemi di reperimento dell'informazione, ma non permette di stimare il contributo dei singoli componenti che formano un sistema di IR. Poiché non è possibile prevedere l'efficacia data da un sistema di IR senza effettivamente testarlo su una collezione sperimentale, risulta necessario valutare un enorme numero di sistemi generati da tutte le possibili combinazioni tra componenti, per permette di confrontare tra loro i vari sistemi e comprendere l'effetto dato da un singolo componente o l'interazione tra questi. In questa tesi si propone uno strumento di Information Visualization per l'esplorazione dei dati di valutazione dei sistemi di IR, chiamato SANKEY. SANKEY aiuta nell'esplorazione delle performance ottenute da un gran numero di sistemi di IR, permettendo di comprendere: quale sistema è il migliore, quali sono i contributi dati da singoli componenti e come questi interagiscono tra loro. Nella versione implementata in questa tesi, SANKEY è stato testato con più di un milione di dati di valutazione la cui esplorazione, senza l'aiuto di un sistema visuale, richiederebbe un'analisi manuale particolarmente onerosa e complessa. Una fase di validazione effettuata attraverso un test di usabilità ha dimostrato come il sistema proposto risulti essere particolarmente efficace e intuitivo da utilizzare.

Indice

1 Introduzione

I sistemi di reperimento dell'informazione hanno lo scopo di recuperare documenti rilevanti a fronte di una esigenza informativa che viene espressa da un utente attraverso una interrogazione composta da una serie di termini descrittivi [Croft et al., 2009]. Tali sistemi sono complessi, poiché caratterizzati da più componenti, come ad esempio stop list, stemmer e modelli, che svolgono funzioni ben precise e tra loro differenti. Queste componenti, combinate insieme, permettono di rappresentare in modo efficiente i documenti della collezione, le query e di recuperare una lista di risultati in grado di soddisfare al meglio l'esigenza espressa da un utente. La valutazione di un sistema di IR è una fase necessaria per comprendere ed analizzare il comportamento del sistema con lo scopo di migliorarne il livello di efficacia necessario per soddisfare le aspettative dell'utente. Un sistema di IR non produce risposte esatte, ma invece classifica i risultati ad una interrogazione dell'utente con una stima di rilevanza, ed è quindi fondamentale riuscire a valutare la qualità dei risultati. Le performance ottenute da diversi sistemi di IR devono essere tra loro comparabili, ed è quindi essenziale che gli esperimenti siano ripetibili attraverso l'utilizzo di collezioni di dati condivise e disponibili pubblicamente.

Il paradigma di Cranfield, largamente utilizzato fin dagli anni '60 nella valutazione dei sistemi di IR, definisce le basi della valutazione in IR attraverso il concetto di collezione sperimentale. Una collezione è definita come una

combinazione di: un insieme di documenti che rappresentano un certo dominio di interesse, un insieme di topic che sintetizzano delle esigenze informative e un insieme di giudizi di rilevanza. La produzione delle collezioni sperimentali è un'attività molto impegnativa che viene effettuata durante campagne internazionali quali TREC (Text REtrieval Conference), CLEF (Conference and Labs of Evaluation Forum), NTCIR (NII Testbeds and Community for Information access Research) e FIRE (Forum for Information Retrieval Evaluation). Grazie a queste collezioni sperimentali è possibile valutare e confrontare i risultati di rilevanza ottenuti da due o più sistemi [Harman, 2011]. Tuttavia, i sistemi di IR vengono attualmente valutati esclusivamente come "*black box*" e non viene valutato l'effetto dato da un singolo componente di un sistema o l'interazione tra componenti. L'unico modo per misurare l'impatto dato dai componenti nelle performance di un sistema di IR è quello di testare tutte le possibili combinazioni di tali componenti, producendo quella che viene definita Grid of Points, una raccolta di dati di valutazione in cui vengono rappresentate tutte le possibili combinazioni di componenti che formano un sistema di IR in modo da permettere di confrontare un gran numero di sistemi ed evidenziare l'effetto generato da un singolo componente o una interazione tra componenti. Per dare un'idea, utilizzando sei stoplist, sei stemmer e diciasette modelli di IR, si creano $6 \cdot 6 \cdot 17 = 612$ diversi sistemi di reperimento. Supponendo che questi vengano testati su sei diverse collezioni sperimentali (ciascuna composta da cinquanta topic) utilizzando sei metriche di valutazione, si hanno $1.101.600$ dati di valutazione La produzione di questi dati richiede enormi sforzi e risorse, inoltre l'esplorazione e l'analisi risulta onerosa e complessa, richiedendo l'utilizzo di complessi strumenti statistici i cui risultati non sono di facile interpretazione per i non statistici, quali ad esempio: *General Linear Mixed Model* (GLMM) [Ferro and Silvello, 2016] e *Analysis Of Variance* (ANOVA) [Ferro and Silvello, 2017].

Utilizzando tecniche di *Information Visualization* (IV) e *Visual Analytics* (VA) è possibile aiutare l'utente a ridurre lo sforzo necessario per l'esplorazione

e l'analisi dei dati della valutazione. Tuttavia, tecniche di IV vengono generalmente utilizzate in IR con lo scopo di presentare ed esplorare i documenti gestiti da un sistema di reperimento dell'informazione e solo in pochi casi per facilitare la valutazione sperimentale.

In questa tesi viene proposto uno strumento visuale, chiamato SANKEY, per l'esplorazione e l'analisi dei dati della valutazione dei sistemi di reperimento dell'informazione. SANKEY, attraverso l'utilizzo di una rappresentazione che prende spunto dai diagrammi di Sankey, permette l'esplorazione dei dati di valutazione di 612 sistemi di IR prodotti dalla combinazione di sei stop list, sei stemmer e diciassette modelli. I 612 sistemi sono valutati utilizzando sei metriche di valutazioni su sei diverse collezioni sperimentali, ciascuna composta da cinquanta topic. In totale si hanno più di un milione di dati di valutazione che richiederebbero una fase di esplorazione manuale troppo onerosa.
Il sistema SANKEY permette all'utente di interagire con i dati:

- selezionando la collezione sperimentale che vuole analizzare;
- decidendo se considerare un singolo topic della collezione o se visualizzare i dati di valutazione per l'intera collezione;
- scegliendo una delle sei metriche di valutazione utilizzate (AP, RBP, P10, nDCG, Twist ed ERR);
- selezionando le componenti da considerare per ciascuna categoria (stoplist, stemmer, modelli).

I dati di valutazione vengono rappresentati attraverso un diagramma di Sankey (figura 1.1) dove i vari nodi rappresentano le componenti di un sistema di IR e gli archi rappresentano interazioni tra questi componenti. Un sistema di IR viene quindi rappresentato da un percorso che congiunge una stoplist, uno stemmer, un modello e un valore finale rappresentante la categoria che identifica la performance ottenuta dal sistema.

Riassumendo, il sistema SANKEY, attraverso una intuitiva interfaccia grafica e numerose funzionalità assiste l'utente nei seguenti compiti:

1. Individuazione dei sistemi di IR migliori (per una determinata collezione sperimentale, uno specifico topic, considerando una particolare metrica di valutazione);

2. Analisi dell'impatto che ha un componente specifico nelle performance di un sistema di IR;

3. Valutazione dell'effetto di interazione tra componenti.

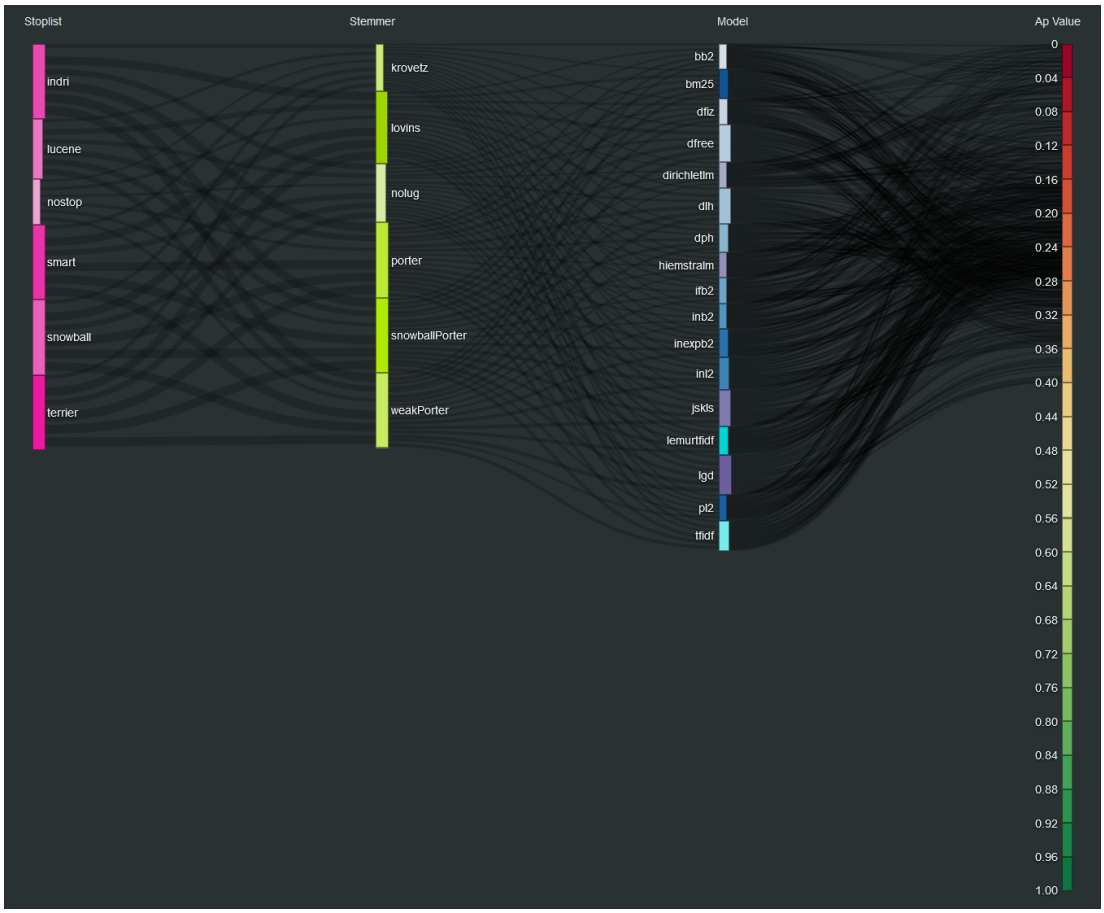


Figura 1.1. *Esempio di rappresentazione dei dati di valutazione attraverso il sistema SANKEY.*

Questo elaborato è organizzato come segue:

- nel capitolo 2 viene introdotto il dominio di applicazione, ovvero il reperimento dell'informazione e le metodologie di valutazione;
- nel capitolo 3 viene descritto il setup sperimentale, ovvero l'insieme di elementi (collezioni sperimentali, Grid of Points e metriche di valutazione) utilizzati dal sistema SANKEY;
- nel capitolo 4 si descrive il sistema SANKEY evidenziandone due componenti principali, ovvero la sezione di selezione dei parametri e l'area di analisi e valutazione dei sistemi di IR. Inoltre si presenteranno degli esempi di utilizzo del sistema;
- nel capitolo 5 viene discussa la validazione del sistema, effettuata attraverso un test di usabilità che ha messo a confronto il sistema SANKEY con un altro strumento visuale chiamato CLAIRE;
- nel capitolo 6 si presentano i related works, in particolare saranno presentati VIRTUE, VATE^2 e CLAIRE;
- nel capitolo 7 vengono discusse le conclusioni e si identificano possibili sviluppi futuri.

2

Il reperimento dell'informazione e la valutazione

Il reperimento dell'informazione o information retrieval (IR) ha lo scopo principale di recuperare informazioni al fine di soddisfare le esigenze informative di un utente.

Il settore del Reperimento dell'Informazione si è inizialmente concentrato sulla gestione di documenti testuali, quali ad esempio libri o articoli digitalizzati. Le informazioni racchiuse all'interno di questi documenti sono solitamente espresse in linguaggio naturale e seguono quindi una forma di espressione non strutturata. Questo è stato il principale motivo per cui l'IR ha assunto un ruolo dominante in alcuni settori rispetto ad esempio alla tradizionale ricerca sulle basi di dati, che agisce su dati strutturati, organizzati secondo schemi e tabelle. Grazie all'avvento di Internet, i motori di ricerca web sono diventati un classico esempio, anche piuttosto potente ed elaborato, di sistemi di IR. Il motore di ricerca web si occupa di recuperare per l'utente, a fronte di una ricerca solitamente composta da poche parole chiave, un insieme di pagine web che dovrebbero soddisfare la sua richiesta di informazioni. Queste vengono ricercate tra miliardi di pagine e vengono presentate all'utente in una lista ordinata in base ad un certo criterio di rilevanza che può tenere conto non soltanto del contenuto della pagina, ma anche di altri fattori, quali ad esempio la notorietà e l'autorevolezza del sito web dal quale la pagina proviene. Grazie ad un sistema di IR, detto anche motore di ricerca, l'utente riesce a reperire le informazioni raccolte in una serie di documenti attraverso

una interrogazione o query, ovvero una raccolta di termini descrittivi per la sua esigenza informativa.

Quindi l'IR si concentra sull'analisi, la rappresentazione, la memorizzazione, l'organizzazione, la distribuzione, l'accesso e il reperimento dell'informazione che può essere contenuta in articoli, e-mail, pagine web o oggetti multimediali quali immagini, video o suoni.

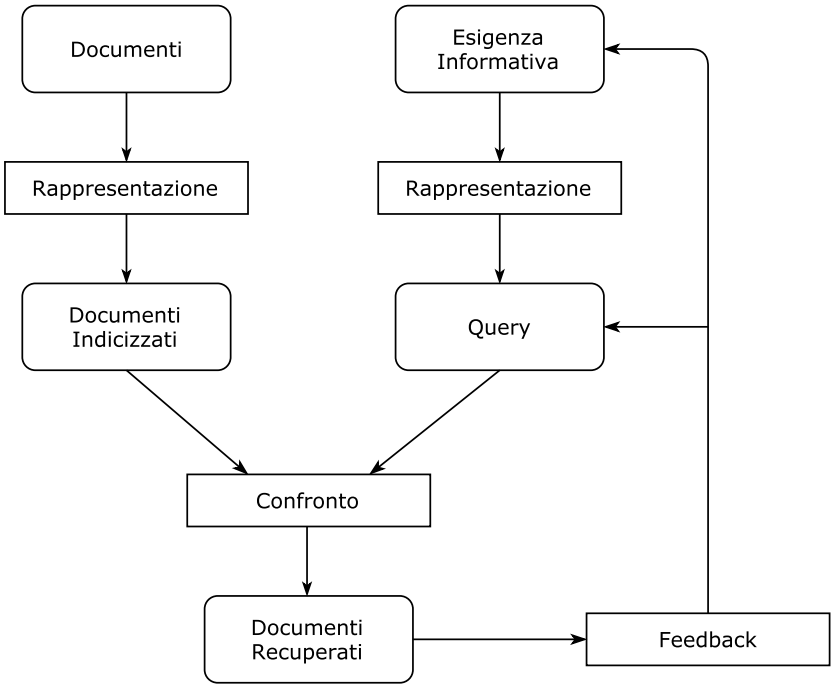


Figura 2.1. *Schema del processo di Reperimento dell'Informazione.*

Qualsiasi contenitore di informazioni viene detto documento, pertanto una collezione di documenti è un insieme da rappresentare, descrivere, memorizzare e da gestire in modo automatico. Un sistema di IR si deve occupare della rappresentazione del contenuto di un documento, della rappresentazione del bisogno informativo dell'utente e per finire del confronto tra le due rappresentazioni al fine di ricavare tale lista ordinata in base al livello di rilevanza (figura 2.1). Quindi i sistemi di IR non operano sui documenti originali, ma su una rappresentazione o vista logica degli stessi che viene ricavata tramite un processo detto di *indicizzazione*.

2.1 Indicizzazione

L'indicizzazione è un processo automatico che ha l'obiettivo di rappresentare il contenuto informativo di un documento. Attraverso appositi algoritmi vengono estratti i *termini indice* per formare un indice che possa permettere di effettuare in modo semplice e veloce il reperimento dell'informazione. La fase di indicizzazione comprende varie fasi (figura 2.2) che verrano qui di seguito descritte. Non è detto che un sistema di reperimento dell'informazione implementi tutte le fasi, ogni fase necessita di calcoli aggiuntivi e non sempre i costi temporali introdotti vengono compensati da un miglioramento significativo nel reperimento finale dei documenti rilevanti. Queste considerazioni sono molto importanti soprattutto per i motori di ricerca web che spesso richiedono il giusto compromesso tra efficacia ed efficienza.

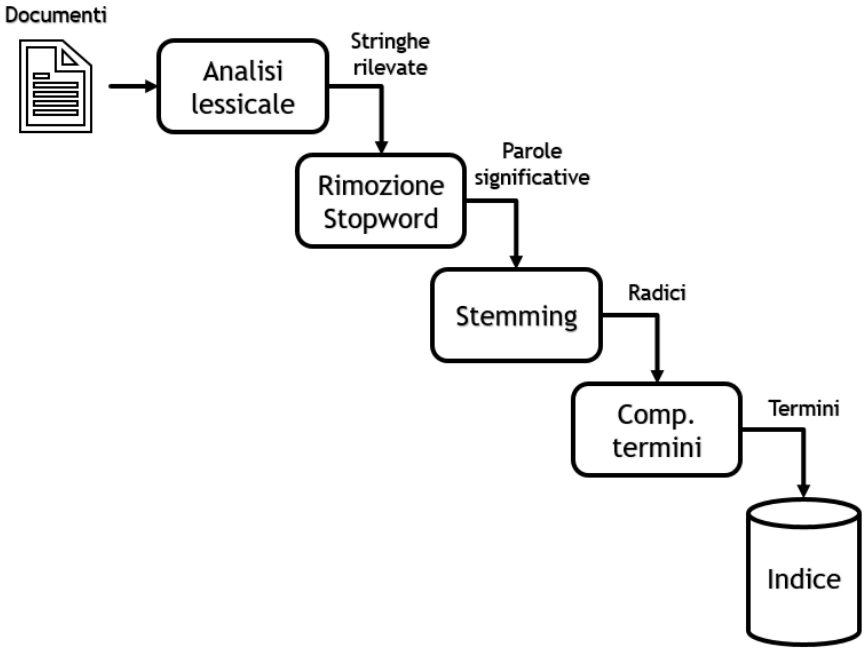


Figura 2.2. *Schematizzazione del processo di indicizzazione.*

2.1.1 Analisi lessicale

Attraverso l'analisi lessicale il testo di un documento viene scansionato per estrarre i token, cioè i potenziali descrittori del documento. I token vengono rilevati attraverso i caratteri di separazione come gli spazi o i segni di punteggiatura. Per questo motivo tale processo dipende fortemente dalla lingua dei documenti, in quanto lingue diverse possono avere alfabeti diversi e caratteri di separazione differenti.

2.1.2 Rimozione delle stopword

Vengono denominate *stopword* le parole funzionali, ovvero quelle parole che hanno scarso contenuto informativo poiché non godono di un significato proprio, ma svolgono una funzione sintattica . Tra queste rientrano solitamente preposizioni, articoli, congiunzioni e pronomi personali. Le stopword differiscono a seconda della lingua utilizzata e vengono raccolte in una lista detta *stop list*. Il processo di rimozione delle stopword permette di ridurre le dimensioni dell'indice finale prodotto dall'intera fase di indicizzazione eliminando proprio le parole che sono considerate di scarso interesse.

2.1.3 Stemming

Lo stemming è la fase dell'indicizzazione che si occupa dell'identificazione della radice (*stem*) di una parola. Attraverso lo stemming è possibile catturare le relazioni esistenti tra le varianti di una parola ottenendo una radice semantica comune, in modo da risolvere delle possibili mancate corrispondenze in fase di confronto tra query e documento. Per molte lingue le parole che provengono dalla stessa radice linguistica hanno una parte o sottostringa comune. Per questo motivo spesso le tecniche di stemming si riconducono alla rimozione di suffissi. Bisogna tuttavia considerare che ciascuna lingua, anche quelle flessive, presentano delle eccezioni che possono condurre un algoritmo di stemming (detto *stemmer*) ad errore. Gli errori si

possono classificare come errori di *over-stemming*, quando una parola viene ricondotta ad uno stem più corto della sua reale radice, o alternativamente di *under-stemming*, quando lo stem risultante è più lungo dello stem corretto per la parola considerata.

2.1.4 Composizione dei termini

Considerando che le interrogazioni poste dall'utente spesso non si limitano ad una sola parola, la composizione dei termini in frasi o gruppi di token è una fase importante dell'indicizzazione. Un documento che contiene esattamente la frase cercata dall'utente è probabilmente più significativo di un documento che la contiene solo parzialmente. Questo risulta particolarmente importante in alcune situazioni, soprattutto perché una composizione, che può essere dei termini o degli stem nel caso sia stata applicata una tecnica di stemming, permette di rendere più specifici termini generici. Alcuni termini associati con altri assumono un significato nuovo e quindi rappresentano una esigenza informativa particolare. Tuttavia la composizione dei termini richiede un'analisi computazionalmente onerosa richiedendo di considerare tutte le diverse composizioni per una query, basandosi sulle diverse definizioni di frase che possono essere utilizzate.

2.1.5 Indice trasposto

L'indice costituisce l'output dell'indicizzazione e ha il compito di memorizzare l'elenco dei termini presenti nei documenti della collezione. Molti dei sistemi di IR moderni utilizzano una struttura dati specifica denominata *indice trasposto* o *inverted index*. Ogni termine indice ha associata una *lista trasposta* (inverted list), detta anche *posting list*, che mantiene tutti i dati relativi al termine considerato. Ogni elemento della lista viene detto *posting* e può contenere differenti informazioni che permettono di determinare il grado di rilevanza di un documento per una determinata query. Ciascun posting, oltre all'identificativo del documento a cui appartiene il termine

indice, può ad esempio contenere la *term frequency* (tf), ovvero il numero di occorrenze del termine nel documento; alternativamente può essere utile considerare la posizione del termine nel documento, in modo da poter determinare possibili relazioni di vicinanza tra termini.

2.2 Modelli di reperimento dell'informazione

A seguito della fase di indicizzazione il sistema di IR deve restituire all'utente in risposta ad una query i documenti recuperati dalla collezione in ordine decrescente di rilevanza. Questo viene effettuato da un algoritmo di reperimento, il quale data un'interrogazione e un documento decide quanto quest'ultimo sia in grado di soddisfare l'esigenza informativa espressa dall'utente. Nel reperimento dell'nformazione esistono diversi *modelli* in grado di definire un insieme di costrutti con lo scopo di formalizzare la rappresentazione dei documenti e la rappresentazione delle interrogazioni. I modelli, inoltre, rendono possibile la realizzazione dell'algoritmo di reperimento dei documenti in risposta ad una query, definiscono cioè il meccanismo che viene utilizzato per comprendere quali documenti della collezione sono rilevanti per una interrogazione. I tre modelli principali che verrano qui di seguito presentati sono il modello booleano, il modello vettoriale e il modello probabilistico.

2.2.1 Modello booleano

Il modello booleano, proposto negli anni '50, viene utilizzato in sistemi industriali, motori di ricerca, biblioteche digitali, OPAC (Online Public Access Catalogue) e sistemi di gestione archivi ed è basato sulla teoria insiemistica e l'algebra di Boole [Croft et al., 2009]. In questo modello i descrittori corrispondono ad insiemi di documenti, mentre le query vengono specificate come espressioni booleane dove gli operandi sono descrittori, mentre gli

operatori sono i classici operatori della logica booleana (AND, OR, NOT). Tale modello, nonostante la sua semplicità, presenta diversi svantaggi:

- le query devono essere espresse sotto forma di espressione booleana e questo richiede un certo livello di addestramento da parte dell'utente;
- la ricerca dei documenti è basata su un criterio di decisione binaria, un documento può essere soltanto rilevante o non-rilevante e non esiste nessun ordinamento per una qualche misura di "similarità";
- non si ha controllo sul numero di documenti recuperati, si rischia di avere nessun documento rilevante quando l'espressione booleana è particolarmente stringente (numero elevato di operatori AND) o viceversa si rischia di avere un insieme risultante troppo grande per poter essere utilizzato in modo efficace (numerosi operatori OR).

Per questi motivi al modello booleano si tende a preferire l'utilizzo del modello vettoriale.

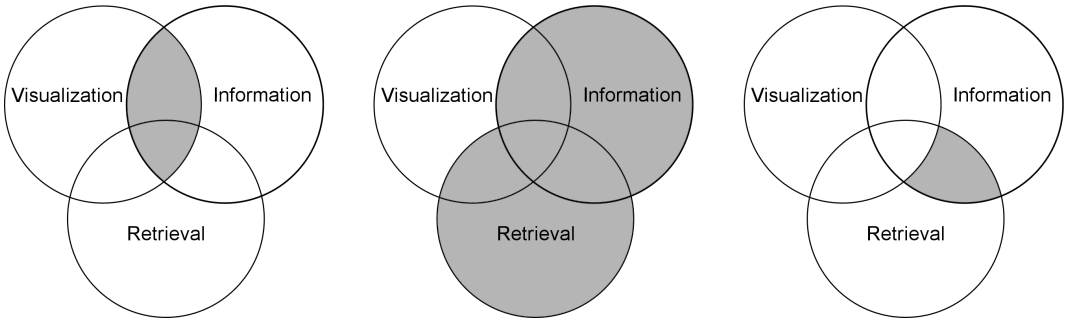


Figura 2.3. *Esempio di combinazioni booleane. Ogni cerchio rappresenta un insieme di documenti. La figura a sinistra mostra i documenti rilevanti per la query "Information AND Visualization", la figura centrale i documenti rilevanti per l'interrogazione "Information OR Retrieval", mentre la figura a destra quelli per la query "(Information AND Retrieval) NOT Visualization".*

2.2.2 Modello vettoriale

Il modello vettoriale è stato proposto da Gerald Salton per risolvere i limiti del modello booleano [Salton et al., 1975]. Tale modello è in realtà un

framework che permette di realizzare una pesatura dei termini non binaria e l'ordinamento dei risultati. I pesi assegnati ai termini indice vengono utilizzati per misurare il grado di similarità tra i documenti della collezione e l'interrogazione dell'utente. I documenti restituiti sono ordinati in maniera decrescente in base al valore di similarità. Formalmente questo modello assume che gli n documenti della collezione e le query appartengano ad uno spazio vettoriale di dimensione t, dove t è il numero dei termini indice. Quindi un generico documento D_i può essere espresso come un vettore

$$D_i = (d_{i1}, d_{i2}, ..., d_{it})$$

dove d_{ij} rappresenta il peso del termine j-esimo nel documento i-esimo. Allo stesso modo la query Q viene espressa come

$$Q = (q_1, q_2, ..., q_t)$$

dove q_j rappresenta il peso del termine j-esimo nella query. Data questa particolare rappresentazione vettoriale i documenti possono essere ordinati utilizzando il risultato del calcolo della distanza fra il vettore di ciascun documento e la query. Questo può essere ricavato ad esempio utilizzando la misura *cosine correlation* che calcola il coseno dell'angolo fra i vettori dei documenti e il vettore della query attraverso la formula:

$$Cosine(D_i, Q) = \frac{\sum_{j=1}^{t}(d_{ij} \cdot q_j)}{\sqrt{\sum_{j=1}^{t} {d_{ij}}^2 \cdot \sum_{j=1}^{t} {q_j}^2}} \tag{2.1}$$

Per il calcolo del grado di similarità è necessario definire uno schema di pesatura per i termini indice. Gran parte dei modelli vettoriali utilizzano uno schema di pesatura del tipo $tf \cdot idf$ dove:

- *tf* è la *term frequency*, ovvero la frequenza del termine in un documento. Questa misura riflette l'importanza del termine in un documento;
- *idf* è la *inverse document frequency*, ovvero l'inverso della frequenza di

un termine nell'interezza dei documenti della collezione. Tale metrica riflette l'importanza di un termine nella collezione.

Lo schema di pesatura $tf \cdot idf$ tiene in considerazione due aspetti fondamentali: se un termine appare spesso in un documento allora è fortemente indicativo circa il contenuto del documento, allo stesso modo però se un termine è presente in tanti documenti non permetterà di distinguerli e quindi sarà poco utile ai fini del reperimento.

2.2.3 Modello Probabilistico

Date le rappresentazioni di query e documento, un sistema presenta un certo grado di incertezza nell'identificazione di un documento rilevante. La teoria della probabilità permette di modellare questo fatto, per questo nel corso degli anni sono stati proposti alcuni modelli probabilistici che si basano sull'idea chiave di stimare quanto probabile sia che un documento risulti rilevante per una data esigenza informativa. L'obiettivo è quello di classificare i documenti in risposta ad una determinata query in ordine di probabilità di rilevanza.

Per fare ciò, data una query Q, per ciascun documento è necessario calcolare $P(R = rilevante|D, Q)$, ossia la probabilità che un documento sia rilevante data l'interrogazione assumendo che tale probabilità dipenda soltanto dalla query e dalla rappresentazione del documento, e ordinare i documenti in base a questo valore. Più la probabilità è alta più il documento considerato è rilevante. Questa è la base del Probability Ranking Principle (PRP) [van Rijsbergen and Jones, 1973].

Insieme al PRP può essere utilizzata una tecnica di reperimento dell'informazione denominata *Binary Independence Model* (BIM) [Robertson and Jones, 1976]. Tale tecnica fa uso delle seguenti assunzioni per rendere possibile la stima della probabilità di somiglianza tra documento e query:

- i documenti sono rappresentati come vettori binari nella forma $\vec{d_j} =$

$(t_1, t_2, t_3, ..., t_N)$ dove t_i assume valore 1 se e solo se il termine i è presente nel documento, 0 altrimenti;

- i termini sono distribuiti in modo indipendente tra i documenti.

Con queste assunzioni, documenti diversi possono essere rappresentati dal medesimo vettore. Anche la query viene rappresentata come un vettore binario di lunghezza pari al numero di termini indice della collezione considerata.

Data una query si può definire un insieme R di risposte ideali per l'interrogazione, che permetta di massimizzare la probabilità di rilevanza. Sia R l'insieme dei documenti rilevanti conosciuti (o inizialmente ipotizzati) per una query q, e $\bar{R}$ il complementare di R, ossia l'insieme dei documenti non rilevanti, si può definire una misura di similarità:

$$sim(d_j, q) = \frac{P(R|\vec{d_j})}{P(\bar{R}|\vec{d_j})} \tag{2.2}$$

dove $P(R|\vec{d_j})$ è la probabilità che il documento d_j sia rilevante per la query q, mentre $P(\bar{R}|\vec{d_j})$ è la probabilità che d_j sia non rilevante per q. Utilizzando la regola di Bayes, la 2.2 può essere riscritta come:

$$sim(d_j, q) = \frac{P(\vec{d_j}|R, q) \times P(R, q))}{P(\vec{d_j}|\bar{R}, q) \times P(\bar{R}, q))} \sim \frac{P(\vec{d_j}|R, q)}{P(\vec{d_j}|\bar{R}, q)} \tag{2.3}$$

dove $P(\vec{d_j}|R, q)$ è la probabilità che qualora venga recuperato un documento rilevante la sua rappresentazione sia uguale a d_j, mentre $P(R, q)$ identifica la probabilità che un documento selezionato in maniera arbitraria dall'intera collezione sia rilevante; il significato di $P(\vec{d_j}|\bar{R}, q)$ e $P(\bar{R}, q)$ è analogo e complementare ai precedenti. Tramite le assuzioni date dal Binary Indipendence Model è possibile stimare queste probabilità tramite l'informazione fornita dai termini indice presenti nei documenti, permettendo di determinare l'espressione utilizzata per il ranking dei documenti:

$$sim(d_j, q) \sim \sum_{t_i \in q \wedge t_i \in d_j} \log\left(\frac{p_{iR}}{1 - p_{iR}}\right) + \log\left(\frac{1 - q_{iR}}{q_{iR}}\right) \tag{2.4}$$

dove è stato posto $p_{iR} = P(t_i|R)$ e $q_{iR} = P(t_i|\bar{R})$ ed è stato assunto $\forall t_i \notin q, p_{iR} = q_{iR}$. La derivazione formale della 2.4 è presentata in dettaglio in [Baeza-Yates and Ribeiro-Neto, 1999]. Tale espressione richiede di calcolare le probabilità $P(t_i|R)$, ovvero la probabilità che il termine indice t_i sia presente in un documento selezionato casualmente nell'insieme dei documenti rilevanti, e $P(t_i|\bar{R})$, con significato analogo alla precedente, ma per i documenti non rilevanti.

Nonostante questo tipo di modello probabilistico abbia il vantaggio di permettere di ordinare i documenti in maniera decrescente in funzione della probabilità di rilevanza, esso presenta alcuni svantaggi: ipotizza l'indipendenza tra i termini indice, inoltre non prende in considerazione la frequenza di occorrenza dei termini nel documento.

2.3 Valutazione dei sistemi di reperimento dell'informazione

Al fine di comprendere quanto bene un sistema si comporti è necessario valutarlo. L'obiettivo della valutazione dei sistemi di reperimento dell'informazione è quello di misurare non solo l'efficienza dei vari sistemi, ma anche e soprattutto la loro efficacia. Un sistema di IR non produce risposte esatte, ma risponde ad una query di un utente proponendo una serie di risultati ordinati secondo una certa stima di rilevanza. È necessario misurare le performance ottenute da una sistema al fine di comprendere come operi effettivamente il sistema e come sia possibile migliorarlo. Le misure delle performance inoltre devono essere comparabili tra i diversi sistemi e per questa ragione è essenziale che gli esperimenti siano ripetibili e soprattutto che i dati utilizzati siano fissati.

2.3.1 Cranfield: le basi della valutazione

La valutazione sperimentale si basa sul paradigma di Cranfield, sviluppato da Cyril Cleverdon, che ha permesso di definire il concetto di collezione sperimentale [Harman, 2011].

Una collezione sperimentale è una terna $C = \{D, T, RJ\}$ dove: D è un insieme di documenti rappresentanti il dominio di interesse; T è un insieme di topic, che sintetizzano le esigenze informative di un utente; RJ è un insieme di giudizi di rilevanza che associano a ciascun topic $t \in T$ i documenti $d \in D$ rilevanti. Di conseguenza la creazione di una collezione di test è un lavoro oneroso, in quanto richiede di definire l'insieme dei documenti che formano la collezione, un insieme di topic che siano in grado di rappresentare accuratamente le esigenze informative degli utenti per quella specifica collezione e infine di definire i giudizi di rilevanza, operazione che può essere eseguita in tre diverse modalità:

- attraverso la valutazione di ogni documento della collezione da parte di un gruppo di esperti, operazione da eseguire per ciascun topic;
- utilizzando un campionamento casuale dei documenti per ciascun topic, metodo che richiede comunque un numero minimo di documenti da considerare molto elevato;
- utilizzando la tecnica di *pooling*, cioè un campionamento basato su diversi esperimenti condotti sulla collezione.

Negli ultimi decenni, sono nate delle apposite campagne per la creazione di collezioni sperimentali quali ad esempio TREC (Text REtrieval Conference[1]) negli Stati Uniti, CLEF (Conference and Labs of Evaluation Forum[2]) in Europa, NTCIR (NII Testbeds and Community for Information access Research[3]) in Giappone e Asia e FIRE (Forum for Information Retrieval Evaluation[4]) in

[1] http://trec.nist.gov/
[2] http://www.clef-campaign.org/
[3] http://research.nii.ac.jp/ntcir/index-en.html
[4] http://fire.irsi.res.in/fire/2016/home

India. Grazie alle campagne di valutazione su larga scala e alle valutazioni sperimentali si ha la possibilità di valutare e confrontare i risultati di rilevanza ottenuti da due o più sistemi per un dato insieme di interrogazioni. È necessario misurare la qualità dei risultati prodotti da ciascun sistema utilizzando delle misure di efficacia basate sui giudizi di rilevanza. Il sistema visuale proposto utilizza diverse metriche per permettere all'utente di confrontare le performance ottenute da tutti i sistemi considerati. Queste verrano viste in dettaglio in sezione 3.3, ma prima sarà necessario introdurre alcuni concetti preliminari.

2.3.2 Definizioni utili

In questa sezione saranno presentate alcune importanti definizioni che rappresentano le fondamenta della valutazione sperimentale nel reperimento dell'informazione [Angelini et al., 2014].

Dato un insieme di giudizi di rilevanza in una collezione sperimentale è essenziale poter associare a ciascun documento, per ogni topic, un grado di rilevanza. Di conseguenza sia *REL* un insieme totalmente ordinato di giudizi o gradi di rilevanza è possibile dare la definizione di *Ground Truth*.

Definizione 2.1. Sia D un insieme finito di documenti e T un insieme finito di topic, la Ground Truth è una funzione definita come:

$$GT : T \times D \rightarrow REL$$
$$(t, d) \mapsto rel$$

Tale funzione associa ad un documento, per un dato topic, il rispettivo valore di rilevanza, permettendo quindi di definire una collezione sperimentale come $C = \{D, T, GT\}$.

A questa segue la definizione di *Recall Base*.

Definizione 2.2. La Recall Base è una funzione:

$$RB : T \to \mathbb{N}$$
$$t \mapsto RB_t = |\{d \in D | GT(t,d) \succ \min(REL)\}|$$

La Recall Base rappresenta il numero totale di documenti rilevanti per un dato topic t, cioè tutti quei documenti che hanno valore di rilevanza superiore al grado "non rilevante" che è il grado minimo dell'insieme REL. È ora possibile definire il concetto di *run*.

Definizione 2.3. Sia T un insieme finito di topic, D un insieme finito di documenti e $N \in \mathbb{N}^+$ la *lunghezza di una run*, una *run* è una funzione:

$$R : T \to D^N$$
$$t \mapsto \mathbf{r}_t = (d_1, d_2, ..., d_N)$$

tale che $\forall t \in T, \forall j,k \in [1,N] | j \neq k \Rightarrow \mathbf{r}_t[j] \neq \mathbf{r}_t[k]$ dove $\mathbf{r}_t[j]$ indica il j-esimo elemento del vettore $\mathbf{r}_t$.

La run per ciascun topic t definisce un vettore di documenti $\mathbf{r}_t$ di lunghezza N, il quale rappresenta una lista ordinata di documenti recuperati per tale topic t con il vincolo che non siano presenti documenti ripetuti.

Per finire vengono introdotte due funzioni che rappresentano la base per il calcolo delle metriche di valutazione, ovvero il *relevance score* e il *relevance weight*.

Il relevance score associa ad ogni elemento di una run il corrispondente grado di rilevanza.

Definizione 2.4. Data una run $R(t) = \mathbf{r}_t$, il relevance score della run è una funzione:

$$\hat{R} : T \times D^N \to REL^N$$
$$(t, \mathbf{r}_t) \mapsto \hat{\mathbf{r}}_t = (rel_1, rel_2, ..., rel_N)$$

dove

$$\hat{\mathbf{r}}_t[j] = GT(t, \mathbf{r}_t[j])$$

Il relevance weight invece associa un intero ad ogni documento per un dato topic, che riflette il grado di rilevanza assegnato dal relevance score.

Definizione 2.5. Sia $W \subset \mathbb{Z}$ un insieme totalmente ordinato e finito di interi, REL un insieme finito di gradi di rilevanza e sia $RW : REL \to W$ una funzione monotona che mappa ogni grado di rilevanza ($rel \in REL$) in un peso di rilevanza ($w \in W$), allora data una run $R(t) = \mathbf{r}_t$, il relevance weight della run è una funzione:

$$\tilde{R} : T \times D^N \to W^N$$
$$(t, \mathbf{r}_t) \mapsto \tilde{\mathbf{r}}_t = (w_1, w_2, ..., w_N)$$

dove

$$\tilde{\mathbf{r}}_t[j] = RW(\hat{\mathbf{r}}_t[j])$$

3 Setup sperimentale

Il sistema proposto permette di visualizzare i dati generati dalla valutazione di una grande raccolta di sistemi di IR. Questi sistemi sono stati valutati su collezioni sperimentali differenti tramite l'utilizzo di diverse metriche di valutazione. In questo capitolo verrano descritte le collezioni di documenti considerate, la Grid of Points (GoP) contenente tutte le combinazioni dei componenti di un sistema di IR scelte, fornita grazie al lavoro presentato in [Ferro and Silvello, 2017] e [Angelini et al., 2017], e per finire si descriveranno in modo formale le metriche di valutazione utilizzate.

3.1 Collezioni di documenti

Sono state considerate le seguenti sei collezioni: TREC 07 Adhoc track, TREC 08 Adhoc track, TREC 09 Web track, TREC 10 Web track, TREC 14 Terabyte track e TREC 15 Terabyte track.

TREC 07 e TREC 08 [Voorhees and Harman, 1998], [Voorhees and Harman, 1999] si concentrano sul task di ricerca delle notizie utilizzando un corpus comprendente 528.155 documenti, ricavati dalle collezioni Text Research Collection Volume 4 (Maggio 1996) e Text Research Collection Volume 5 (Aprile 1997) dalle quali sono stati esclusi i documenti del Congressional Record (1993). Entrambe sono caratterizzate da 50 topic differenti, numerati da 351 a 400 per TREC 7 e da 401 a 450 per TREC 8. I giudizi di rilevanza

associati sono di tipo binario, quindi sia dato un topic $t \in T$ e un documento $d \in D$ questo può essere considerato rilevante o non rilevante.

TREC 09 e TREC 10 [Voorhees and Harman, 2000], [Voorhees, 2001] si focalizzano su task di ricerca web utilizzando un corpus noto con il nome WT10g di 1.692.096 documenti e di dimensione pari a circa 10 gigabyte. Esso comprende un sottoinsieme di pagine web ricavate da una immagine del web del 1997 estratta dalla libreria digitale Internet Archive. Anche in questo caso entrambe le collezioni sono accompagnate da 50 topic numerati rispettivamenti da 451 a 500 e da 501 a 550. Vengono utilizzati giudizi a tre gradi di rilevanza ed in particolare un documento può essere considerato non rilevante, rilevante o molto rilevante (highly relevant).

Anche per TREC 14 e TREC 15 [Voorhees, 2005], [Voorhees, 2006] viene considerato un task di ricerca web. I dati sono estratti da siti web del dominio *.gov* all'inizio del 2004. La collezione, nota con il nome *"GOV2"*, raccoglie 25 milioni di documenti per una dimensione totale di circa 426 gigabyte. I topic associati sono numerati da 751 a 800 per TREC 14 e da 801 a 850 per TREC 15. Proprio come per TREC 09 e TREC 10 si utilizzano giudizi a tre gradi di rilevanza.

3.2 Grid of Points

La Grid of Points[1] consiste in una raccolta di dati nati da una serie di esperimenti in cui vengono idealmente rappresentate tutte le possibili combinazioni di componenti che formano un sistema di reperimento dell'informazione. La Grid of Points utilizzata è stata generata dal lavoro presentato in [Ferro and Silvello, 2017] e [Angelini et al., 2017] considerando tre componenti principali di un sistema di reperimento dell'informazione: stop list, stemmer e il modello di reperimento. In particolare i singoli componenti scelti sono:

- **Stop list**: *nostop*, *indri*, *lucene*, *snowball*, *smart* e *terrier*;

[1] http://gridofpoints.dei.unipd.it/

- **Stemmer**: *nolug*, *weakPorter*, *porter*, *snowballPorter*, *krovetz* e *lovins*;
- **Modelli**: *bb2*, *bm25*, *dfiz*, *dfree*, *dirichletlm*, *dlh*, *dph*, *hiemstralm*, *ifb2*, *inb2*, *inl2*, *inexpb2*, *jskls*, *lemurtfidf*, *lgd*, *pl2*, *tfidf*.

Ogni sistema è caratterizzato dall'utilizzo di una stop list, di uno stemmer e di un modello. I componenti scelti generano un totale di $6 \cdot 6 \cdot 17 = 612$ sistemi, uno per ciascuna combinazione. Tutti questi sistemi sono valutati utilizzando sei diverse metriche (Average Precision, Precision at 10, Normalized Discounted Cumulative Gain, Rank Biased Precision, Expected Reciprocal Rank e Twist) sulle sei diverse collezioni sperimentali precedentemente illustrate (ciascuna accompagnata da 50 topic). In totale si hanno più di un milione di punti, rappresentati dati di valutazione di sistemi di IR, che forniscono un'idea sull'importanza dello sviluppo di un sistema visuale per la rappresentazione e l'esplorazione di questi dati.

3.2.1 Stoplist

Le diverse stop list si differenziano per il numero di stopword che contengono. Oltre al caso *no stop*, dove non si utilizza alcuna stoplist, sono state considerate le stoplist *indri* formata da 418 termini, *lucene* con 33 termini, *snowball* con 174 termini, *smart* con 571 termini e per finire *terrier* contenente 733 termini.

3.2.2 Stemmer

Sono state scelte le tecniche di stemming più utilizzate e che si sono imposte come metodi standard per la lingua Inglese. *Lovins* è stato il primo stemmer disponibile [Lovins, 1968]. Utilizza un dizionario di suffissi comuni insieme ad una lista di eccezioni relative ai suffissi. L'algoritmo richiede due step per determinare la rimozione di un suffisso; data una parola si controlla se questa termina con uno dei suffissi segnalati nel dizionario e, in caso positivo, si verifica se tale parola costituisce un'eccezione per quel suffisso. Se la parola

non rappresenta un'eccezione allora il suffisso può essere rimosso. Lovins è uno stemmer molto aggressivo, poiché il suo dizionario comprende 294 suffissi.

Porter [Porter, 1980] applica lo stesso approccio dello stemmer Lovins, ma è meno agressivo in quanto utilizza un dizionario di circa 60 suffissi. Questo è stato possibile impiegando una tecnica di rimozione dei suffissi ricorsiva. Lo svantaggio è che tale algoritmo risulta più lento dell'algoritmo di Lovins, utilizzando 8 step per generare lo stem di una parola. Per questa tecnica di stemming sono state scelte anche due varianti, *snowballPorter* e *weakPorter*.

La tecnica di stemming *Krovetz* [Krovetz, 1993] sfrutta un'analisi della flessione e della derivazione delle parole che permette di produrre stem morfologicamente corretti. Inizialmente effettua una rimozione dei suffissi trasformando le parole dalla forma plurale al singolare, eliminando il suffisso "ing" e convertendo verbi dal tempo passato al presente. Successivamente il risultato viene confrontato con un dizionario che permette di trasformare lo stem generato dall'iniziale rimozione del suffisso in una parola realmente valida e comprensibile. Rispetto a Porter e Lovins, Krovetz è molto meno aggressivo e cerca di migliorare la precisione e la robustezza correggendo errori di spelling e stem privi di significato.

3.2.3 Modelli

I modelli considerati per la composizione della GoP si differenziano in tre grandi famiglie: i modelli vettoriali, i modelli probabilistici e i language models (figura 3.1).

Nella prima famiglia rientrano i modelli *TFIDF* e la variante *LemurTFIDF* implementata in Lemur. Questi modelli utilizzano la statistica $tf \cdot idf$ (term frequency - inverse document frequency) per stabilire l'importanza di una parola in un documento.

A differenza dei due modelli precedenti, i modelli probabilistici utilizzano una formula motivata dalla teoria della probabilità. In questa tipologia di modelli rietrano il modello BM25 e la famiglia di modelli *Divergence*

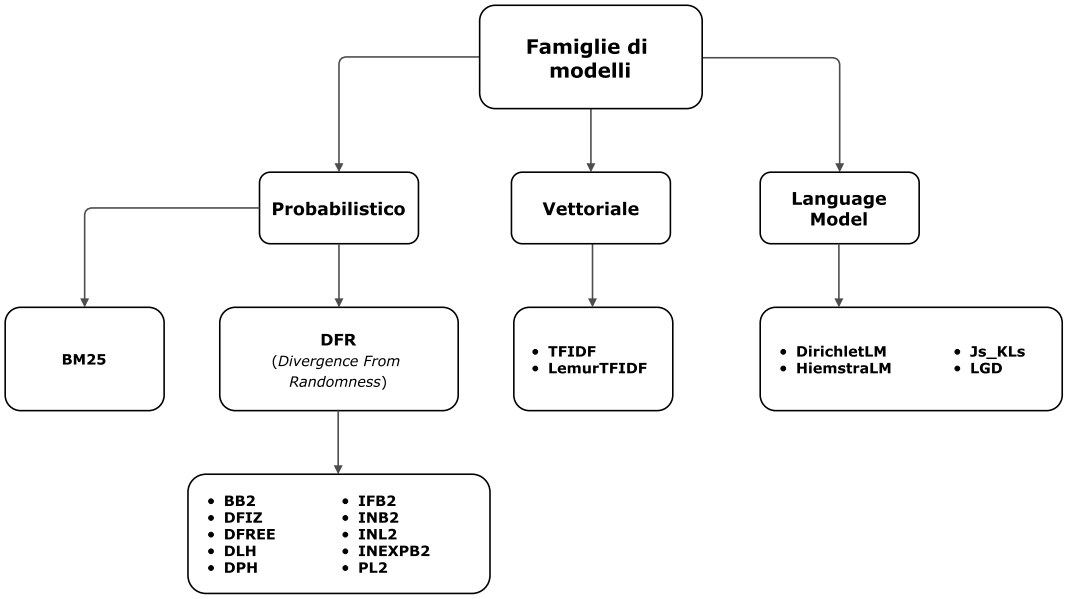


Figura 3.1. *Tassonomia dei modelli utilizzati per la Grid of Points.*

From Randomness (DFR). Il modello *BM25* (Best Match 25), definito da Stephen Robertson e Stephen Walker, si differenzia dal più semplice Binary Indipendence Model presentato in sezione 2.2.3 poiché tiene in considerazione la frequenza dei termini in ciascun documento, la lunghezza del documento e la frequenza dei termini nella query. I modelli DFR si basano sull'idea che se un termine "raro" ha molte occorrenze in un documento allora con grande probabilità è informativo circa l'argomento descritto dal documento. In particolare i pesi del termini vengono calcolati misurando la divergenza tra la distribuzione dei termini prodotta da un processo aleatorio e la distribuzione effettiva dei termini. Questi modelli sono caratterizzati dalla scelta di un processo stocastico e da due tipi di normalizzazione. Una normalizzazione di primo livello che presuppone che i documenti abbiano uguale lunghezza e misura il guadagno informativo ottenuto da un termine una volta che questo sia stato accettato come buon descrittore del documento osservato.Una normalizzazione di secondo livello che è invece legata alla lunghezza del documento e ad altre statistiche [Amati and Van Rijsbergen, 2002]. Questi due metodi di normalizzazione vengono applicati successivamente al modelli di base per ottenere la formula di pesatura dei termini. In questa particolare

Modello DFR	Descrizione
BB2	Distribuzione Bose-Einstein + rapporto di due processi di Bernoulli per la normalizzazione
DFIZ	Modello Divergence From Independence basato su Standardization
DFREE	Distribuzione Ipergeometrica basata su una media di due misure informative
DLH	Distribuzione Ipergeometrica, ma senza parametri
DPH	Distribuzione Ipergeometrica + Normalizzazione di Popper
IFB2	Modello Inverse Term Frequency + rapporto di due processi di Bernoulli per la normalizzazione
INB2	Modello Inverse Document Frequency + rapporto di due processi di Bernoulli per la normalizzazione
INL2	Modello Inverse Document Frequency + successione di Laplace per la normalizzazione
INEXPB2	Modello Inverse Expected Document Frequency + rapporto di due processi di Bernoulli per la normalizzazione
PL2	Distribuzione di Poisson + successione di Laplace per la normalizzazione

Tabella 3.1. *Breve descrizione dei modelli DFR utilizzati nella Grid of Points.*

categoria rientrano i modelli *BB2*, *DFIZ*, *DFREE*, *DLH*, *DPH*, *IFB2*, *INB2*, *INL2*, *INEXPB2* e *PL2*. Una breve descrizione per ciascuno di questi viene fornita nella tabella 3.1.

I language models basano il loro funzionamento sull'idea che un documento sia rilevante per una data interrogazione se il documento contiene spesso i termini che compongono la query [Manning et al., 2008]. Ciascun documento viene visto come un campione della lingua definendo un "modello di documento", mentre una query viene trattata come un processo di generazione. Per ciascun modello di documento si calcola la probabilità di generazione, ovvero la probabilità che l'interrogazione sia generata dal modello considerato. Maggiore è la probabilità di generazione

di un documento più il documento è rilevante per l'interrogazione. I language model che sono stati considerati nella Grid of Points sono DirichletLM (un language model con smoothing bayesiano e Dirichlet prior), HiemstraLM (modello presentato da Djoerd Hiemstra che utilizza una nuova interpretazione probabilistica dello schema di pesatura $tf \cdot idf$ [Hiemstra, 1998]), Js_KLs (un modello che nasce dal prodotto di due misure, la divergenza di Jeffrey e la divergenza di Kullback-Leibler) e LGD (modello che applica lo smoothing di Jelinek-Merce).

3.3 Metriche di valutazione utilizzate

Inizialmente, durante gli studi di Cranfield, vennero introdotte due metriche per la valutazione: Precisione e Richiamo. Per una determinata query precisione e richiamo sono definite rispettivamente come:

$$Precisione = \frac{|Rel \cap Ret|}{|Ret|} \tag{3.1}$$

$$Richiamo = \frac{|Rel \cap Ret|}{|Rel|} \tag{3.2}$$

dove l'operatore $|.|$ ritorna la cardinalità dell'insieme considerato, Rel rappresenta l'insieme di documenti rilevanti (e quindi $|Rel|$ identifica la Recall Base) e Ret rappresenta l'insieme di documenti recuperati dalla run (quindi $|Ret|$ non è altro che la lunghezza della run). Precisione e Richiamo sono la base della valutazione per i sistemi di reperimento dell'informazione, ma non sono misure *ordinate*, ovvero sono calcolate senza tenere il considerazione l'effettivo posizionamento di un documento nella run (rank). Pensando al funzionamento di un motore di ricerca l'utente spesso limita la propria visualizzazione solo alle prime pagine web recuperate, non è interessato a scorrere tutti i risultati che gli sono stati forniti. Di conseguenza un motore di ricerca che produce un documento rilevante nelle prime posizioni della lista proposta all'utente può essere considerato migliore di un sistema che invece

produce tale documento in posizioni più basse della lista, preceduto da un insieme di documenti che non risultano attinenti all'interrogazione dell'utente. Nel corso degli anni sono state proposte metriche di valutazione più potenti, che tengono conto anche di questi aspetti. Il sistema visuale presentato valuta la Grid of Points utilizzando sei diverse metriche di valutazione: *Average Precision*, *Precision at 10*, *Rank-Biased Precision*, *Normalized Discounted Cumulative Gain*, *Expected Reciprocal Rank* e *Twist* che verranno qui di seguito presentate. Tutte le metriche vengono applicate per analizzare le performance di un sistema per un singolo topic della collezione, ma sono presenti anche le controparti che analizzano le performance sull'intera collezione attraverso una media aritmetica dei valori ottenuti su tutti i topic. Di queste verrà presentata a titolo esemplificativo soltanto la Mean Average Precision che è diventata uno standard ed è una delle metriche più utilizzate per misurare le performance di un sistema di reperimento.

3.3.1 Precisione a livello di cut-off K

Come spiegato precedentemente, in molte situazioni è importante valutare un sistema in base ai documenti recuperati e presentati nelle prime posizioni della run considerata. Per questo motivo è stata introdotta la precisione a diversi livelli di cut-off. Tale metrica viene chiamata *Precision at k* [Manning et al., 2008] dove k rappresenta il valore di cut-off. Di solito viene scelto $k = 10$ (*Precision at 10*). Nonostante questa metrica risulti utile per confrontare quale sistema presenti un numero maggiore di documenti rilevanti nelle prime k posizioni, non viene considerata l'effettiva posizione dei documenti rilevanti. Inoltre tale metrica risulta poco significativa per topic con numero di documenti rilevanti minore di k. Supponendo di considerare la precision at k con $k = 10$ e di dover valutare tale metrica per un topic con cinque documenti rilevanti ($RB_t = 5$), un sistema "perfetto" che posiziona tali documenti nelle prime cinque posizioni avrà $Precision\,at\,10 = 0.5$, in contrasto con il valore massimo che è invece pari a 1.

3.3.2 Average Precision e Mean Average Precision

La *Average Precision* (AP) [Manning et al., 2008] è stata introdotta in TREC-2 nel 1993 ed è diventata una delle metriche più utilizzate in IR. Tale metrica rappresenta la media del valore di precisione ottenuto dopo che ogni documento rilevante è stato recuperato. Dato un topic $t \in T$, sia $\{d_1, d_2, ..., d_n\}$ la run di lunghezza n, ovvero l'insieme ordinato dei documenti recuperati, la Average Precision può essere espressa come:

$$AP = \frac{1}{RB_t} \sum_{k=1}^{n} \left(\frac{\tilde{r}_t[k]}{k} \sum_{j=1}^{k} \tilde{r}_t[j] \right) \tag{3.3}$$

dove RB_t rappresenta la recall base per il topic t, mentre $\tilde{r}_t[k]$ rappresenta il relevance weight assegnato al k-esimo documento della run che assume valore pari a 1 se il documento è rilevante, 0 altrimenti.

L'utilizzo di tale metrica presenta diversi vantaggi. È sempre compresa tra 0 e 1 ed è una misura top-heavy, ovvero dà peso maggiore ai documenti presenti nelle posizioni alte del rank. Un documento non rilevante presente nelle prime posizioni influenza più negativamente tale misura rispetto ad un errore nelle posizioni più basse, che viene tuttavia comunque considerato. L'Average Precision è particolarmente adatta per valutare il compito di recuperare il maggior numero di documenti rilevanti possibile, tenendo presente che i documenti recuperati nelle prime posizioni del rank sono i più importanti.

Spesso è necessario valutare un sistema non per uno specifica esigenza informativa, ma per l'intera collezione considerata. In questo caso è possibile calcolare la media aritmetica dell'Average Precision. La media viene calcolata considerando tutti i topic della collezione ricavando quella che viene chiamata Mean Average Precision (MAP) [Manning et al., 2008]. Sia T l'insieme dei topic per una data collezione di documenti, la Mean Average Precision può essere espressa come

$$MAP = \frac{1}{|T|} \sum_{j=1}^{|T|} AP_j \tag{3.4}$$

dove AP_j rappresenta la Average Precision ottenuta dal j-esimo topic della collezione.

3.3.3 Normalized Discounted Cumulative Gain

Tutte le metriche fin qui considerate fanno riferimento a sistemi con solo due gradi di rilevanza, dove i documenti vengono classificati rilevanti o non rilevanti. Tuttavia in diverse situazioni è necessario esprimere più di due gradi di rilevanza e in questi casi si parla di *multigraded relevance* (rilevanza a più gradi). Ad esempio è possibile avere quattro gradi di rilevanza, ciascuno con il proprio *relevance weight* o peso assegnato. Un documento può essere considerato *Highly Relevant*, *Fairly Relevant*, *Partially Relevant* o *Not Relevant*, con peso assegnato di valore descrescente (ad esempio 3, 2, 1, 0). Il *Cumulative Gain* è una metrica utilizzata per calcolare l'efficacia di un sistema quando sono presenti giudizi di rilevanza multigrado, sfruttando i pesi assegnati ai vari gradi di rilevanza.

Sia N il numero di documenti recuperati da un sistema per un dato topic t e sia $j \in N^+$ tale che $1 \leq j \leq N$, il cumulative gain a livello j è definito come

$$CG(j) = \sum_{k=1}^{j} \tilde{r}_t[k] \tag{3.5}$$

dove $\tilde{r}_t[k]$ rappresenta il peso associato al documento posto in posizione k nella run per il topic t.

Anche il Cumulative Gain non tiene espressamente conto della posizione dei documenti nella run. Idealmente trovare un documento Highly Relevant (ovvero di score massimo) nelle prime posizioni dovrebbe essere considerato maggiormente importante rispetto a trovare lo stesso documento nelle posizioni inferiori. Per questa ragione è stata introdotto il *Discounted Cumulative Gain* che grazie all'utilizzo di una *discounting function* riduce progressivamente il peso dei documenti man mano che si procede nel ranking. Sia data dg, la funzione di discounting scelta, N il numero di documenti

recuperati per il topic considerato e $j \in N^+$ tale che $1 \leq j \leq N$, il discounted cumulative gain a livello j viene definito come

$$DCG(j) = \sum_{k=1}^{j} dg(k) \tag{3.6}$$

dove $dg(k)$ rappresenta il nuovo peso, calcolato utilizzando la discounting function, associato al documento in k-esima posizione nella run. Anche questa metrica ha il vantaggio di essere top-heavy, inoltre può essere adattata al task considerato utilizzando una differente funzione di discounting. Tuttavia è una misura non compresa tra i valori 0 e 1 ed i valori che ottiene sono fortemente dipendenti dal topic considerato in quanto il valore massimo dipende dalla specifica esigenza informativa. Per la risoluzione di questo problema è possibile effettuare una normalizzazione utilizzando il concetto di *sistema ideale*. Il sistema ideale rappresenta lo scenario perfetto in cui tutti i documenti rilevanti sono situati nelle posizioni alte del ranking, in ordine decrescente di *relevance score*. Di conseguenza è preferibile utilizzare il *Discounted Cumulative Gain Normalizzato* (nDCG) che viene definito dalla seguente equazione

$$nDCG(j) = \frac{DCG(j)}{iDCG(j)} \tag{3.7}$$

dove il numeratore rappresenta il discounted cumulative gain a livello j del sistema considerato, mentre il denominatore quello del sistema ideale. In questo caso il valore è sempre compreso nell'intervallo $[0, 1]$, rendendo la metrica confrontabile tra topic differenti e permettendo di effettuare una media aritmetica tra tutti i topic per la valutazione di un sistema di IR per un'intera collezione.

3.3.4 Rank-Biased Precision

Un'altra metrica utilizzata è la Rank-Biased Precision (RBP) che, differentemente dall'Average Precision, ha lo scopo di misurare l'utilità che un utente può guadagnare esplorando il sistema. Normalmente un utente

non è interessato ad esplorare tutti i risultati, ma partendo dal primo che gli viene proposto, passa ai successivi scorrendo con un certo grado di persistenza. È possibile identificare due classi di utente: un utente paziente, con un alto valore di persistenza p e che è più propenso ad avanzare nel ranking dei documenti e un utente impaziente che è caratterizzato da un basso valore p e che quindi con scarsa probabilità tenderà a proseguire nella lettura dei documenti proposti. Supponendo che ogni decisione di un utente di proseguire nell'analisi di un run sia indipendente dalla profondità del ranking al quale è giunto, dalla decisione precedente di proseguire o meno nella lista e dal valore di rilevanza del documento che sta esaminando, la Rank-Biased Precision può essere definita come

$$RBP = (1-p) \cdot \sum_{i=1}^{d} r_i \cdot p^{i-1} \tag{3.8}$$

dove r_i è il relevance weight assegnato al documento in i-esima posizione nella run e p^{i-1} rappresenta la probabilità che l'utente arrivi ad osservare l'i-esimo documento. Oltre ad aver il vantaggio di essere una metrica derivata da un modello basato sul comportamento degli utenti, non richiede la conoscenza dell'ampiezza della collezione o il numero di documenti rilevanti per ciascuna query. Presenta invece lo svantaggio, rispetto alla nDCG, di essere una metrica pensata per giudizi di rilevanza binari.

3.3.5 Expected Reciprocal Rank

Come è stato spiegato in 3.3.3, la misura DCG (così come la sua versione normalizzata) è particolarmente adatta in casi di giudizi a più gradi di rilevanza, tuttavia presenta una forma di indipendenza dei dati che in alcuni casi non è desiderata, ad esempio un documento in una data posizione ottiene sempre lo stesso guadagno o la stessa riduzione (a seconda del suo grado di rilevanza) indipendentemente dai documenti mostrati in precedenza. L'Expected Reciprocal Rank (ERR) è una metrica utilizzata per giudizi a più gradi di rilevanza che cerca di risolvere questo problema penalizzando

documenti che sono visualizzati in seguito a documenti molto rilevanti, poiché ci si aspetta che un utente con alta probabilità fermerà la sua ricerca dopo aver trovato un risultato estremamente rilevante. Più precisamente ERR è definita come la durata prevista che impiegherà l'utente per trovare un documento rilevante. L'Expected Reciprocal Rank può quindi essere espressa come

$$ERR = \sum_{r=1}^{n} \frac{1}{r} \cdot P_{stop}(r) \tag{3.9}$$

dove n rappresenta la lunghezza della run, mentre $P_{stop}(r)$ la probabilità che l'utente si fermi alla posizione r. Anche in questo caso, come per RBP, l'utente visualizza sequenzialmente i risultati di ranking dalla prima posizione (che viene sempre visualizzata) fino all'ultimo documento della run. Con probabilità R si ferma e non passa al documento seguente, oppure con probabilità $1 - R$ continua a proseguire passando al rank successivo. R in questo caso dipende dalla rilevanza del documento che l'utente sta analizzando. Sia quindi g_i il grado di rilevanza dell'i-esimo documento, allora

$$R_i = \mathcal{R}(g_i) \tag{3.10}$$

dove $\mathcal{R}$ è una funzione che mappa il grado di rilevanza nella probabilità di rilevanza. Di conseguenza l'equazione (3.9) può essere riscritta come

$$ERR = \sum_{r=1}^{n} \frac{1}{r} \cdot \prod_{i=1}^{r-1} (1 - R_i) \cdot R_r \tag{3.11}$$

ERR è quindi una misura che considera anche i documenti già visualizzati, al contrario di RBP, ed è particolarmente top-heavy poiché penalizza i sistemi che posizionano documenti non rilevanti nelle prime posizioni del rank.

3.3.6 Twist

Twist (τ) è una metrica, valida sia per giudizi di rilevanza binari, sia per giudizi a più gradi di rilevanza, che cerca di quantificare lo sforzo causato

all'utente nella ricerca di documenti rilevanti nella run proposta rispetto alla run ideale [Ferro et al., 2016].
Data una run r_t per uno specifico topic $t \in T$, Twist può essere espressa come

$$\tau_{r_t} = \frac{\rho_{r_t} + \sigma_{r_t}}{2} \tag{3.12}$$

dove ρ_{r_t} è il *recovery ratio* che stima quanto vicino è il *balance point* rispetto alla recall base, σ_{r_t} è lo *space ratio*, una media armonica tra forward e backward space ratio, che misura l'area sotto la curva *Relative Position* (RP curve).

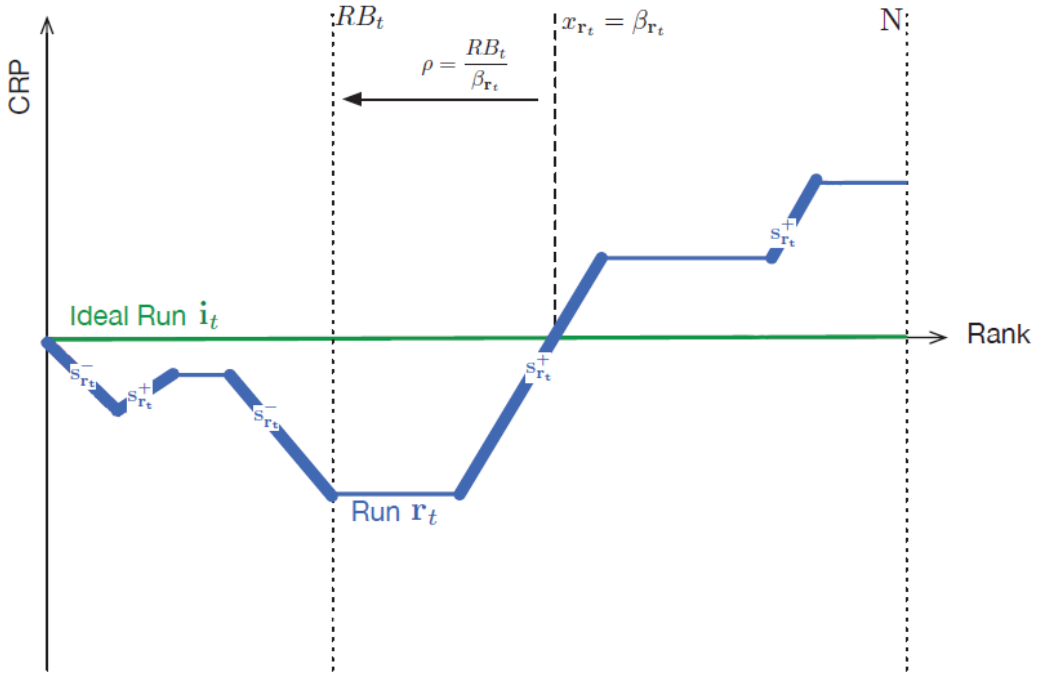


Figura 3.2. *Visualizzazione intuitiva di recovery e space ratio [Ferro et al., 2016].*

La Relative Position valuta quanto un documento è situato in posizione errata nella lista ordinata proposta rispetto al caso ideale, quindi un valore zero individua un documento in posizione ottima, un valore maggiore di zero un documento di alta rilevanza situato in una posizione dove era atteso un documento di rilevanza inferiore, infine un valore minore di zero individua un documento posto in posizioni più elevate nella run rispetto al caso ideale. È possibile definire anche la *Cumulated Relative Position* (CRP), la quale per ogni

posizione del ranking, considera la somma dei valori di RP fino alla posizione considerata.

In figura 3.2 vengono illustrati i concetti di space ratio e recovery ratio attraverso l'utilizzo di una curva CRP di una run confrontata con il caso ideale. La recovery ratio misura quanto vicino alla recall base è situato il punto di attraversamento della curva CRP all'asse x, più vicino è il punto migliore è la run. Lo space ratio misura l'area sotto la curva RP, ovvero la lunghezza della curva CRP sia per errati posizionamenti nel ranking con effetto positivo che per errati posizionamenti con effetto negativo (evidenziati con spessore maggiore nella curva), minore è il valore migliore è la run poiché presenta meno documenti posti in posizioni sbagliate rispetto alla run ideale.

4 Il sistema SANKEY

In questo capitolo presentiamo lo strumento di visualizzazione SANKEY (disponibile al sito: http://gridofpoints.dei.unipd.it/sankey). Il sistema visualizza i dati utilizzando una rappresentazione grafica che prende spunto dal *diagramma di Sankey*. Il diagramma di Sankey è una tecnica di visualizzazione generalmente utilizzata per rappresentare un flusso da un insieme di valori ad un altro. Nodi che rappresentano insiemi categorici vengono connessi tra loro attraverso dei link o archi che assumono un dimensionamento in larghezza proporzionale alla quantità di flusso. Per queste ragioni i diagrammi di Sankey vengono generalmente applicati per evidenziare trasferimenti di energia, materiali, costi o dati in un processo. Vedremo come questa tecnica di visualizzazione verrà adattata per permettere l'esplorazione e l'analisi dei dati di valutazione dei sistemi di reperimento dell'informazione. La figura 4.1 permette di avere una visione complessiva del sistema SANKEY, la quale permette di evidenziare due aree principali: una sezione di selezione dei parametri e la sezione in cui vengono rappresentati i dati di valutazione.

4.1 Descrizione del sistema

Il sistema SANKEY si suddivide principalmente in due sezioni:

1. La sezione di *selezione dei parametri* (figura 4.2) che permette all'utente

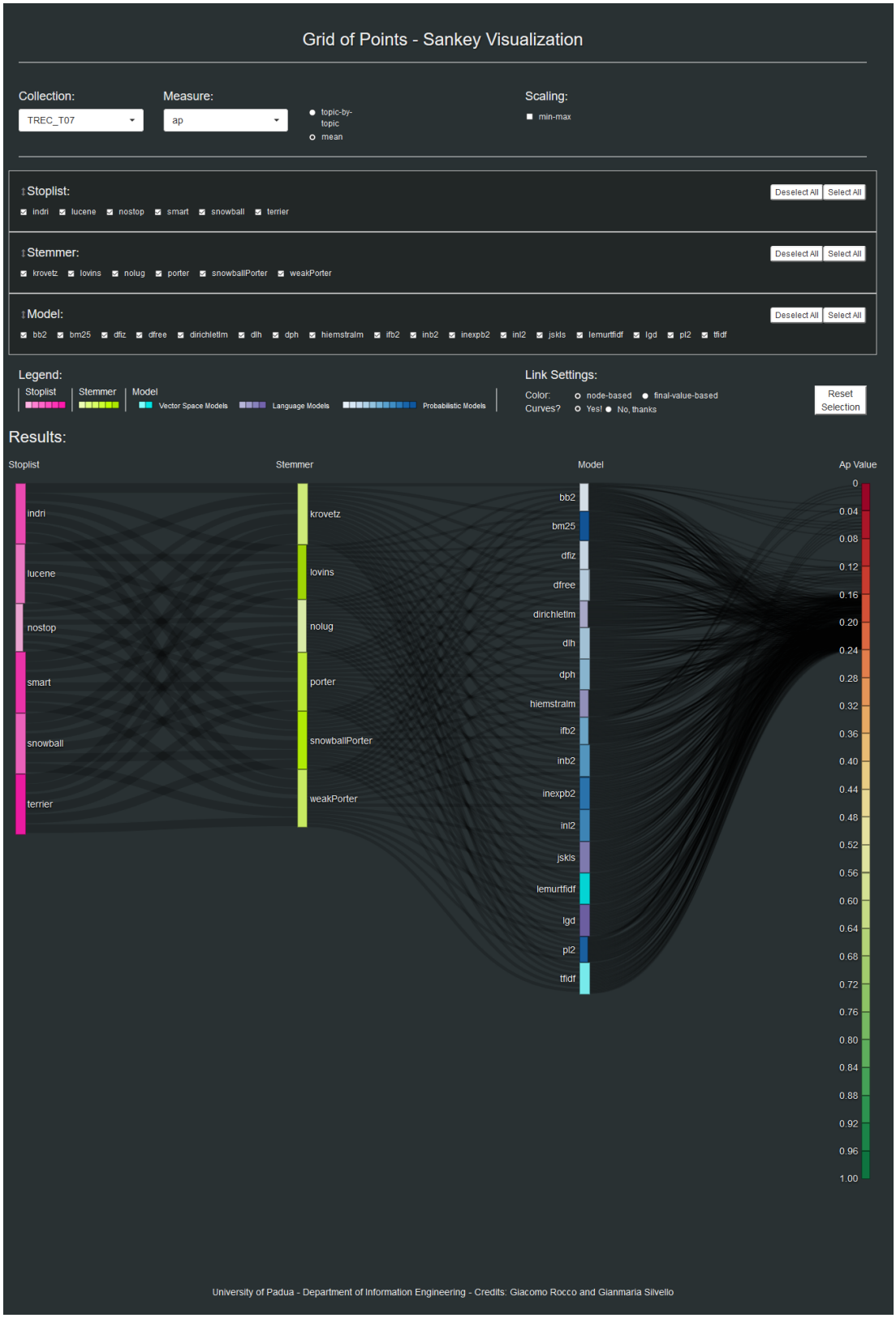


Figura 4.1. *Visualizzazione complessiva del sistema SANKEY.*

di caricare i dati per il sistema, ad esempio scegliendo la collezione di documenti da analizzare, l'insieme di componenti da visualizzare e la metrica di valutazione da considerare.

2. L'area di *analisi e valutazione dei sistemi di IR* (figura 4.3) situata dopo la sezione di selezione dei parametri, che permette l'effettiva esplorazione dei sistemi di IR e delle rispettive valutazioni in base alle configurazioni scelte nell'area di selezione dei parametri.

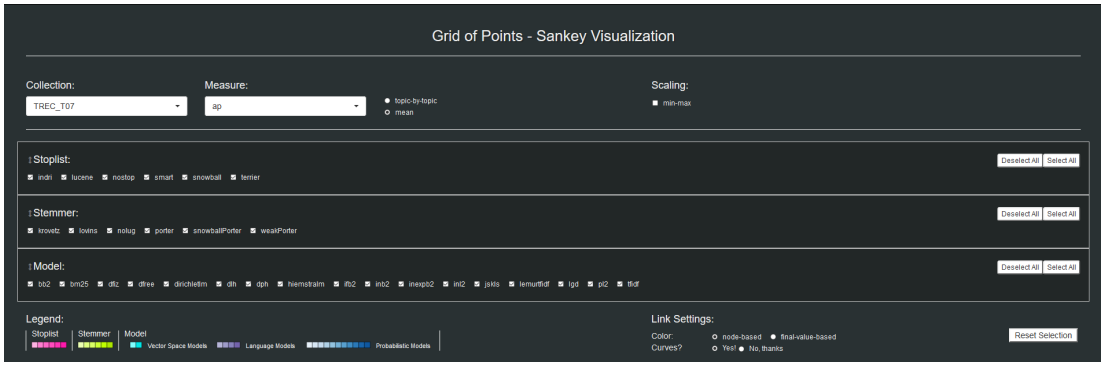


Figura 4.2. *Area di selezione dei parametri.*

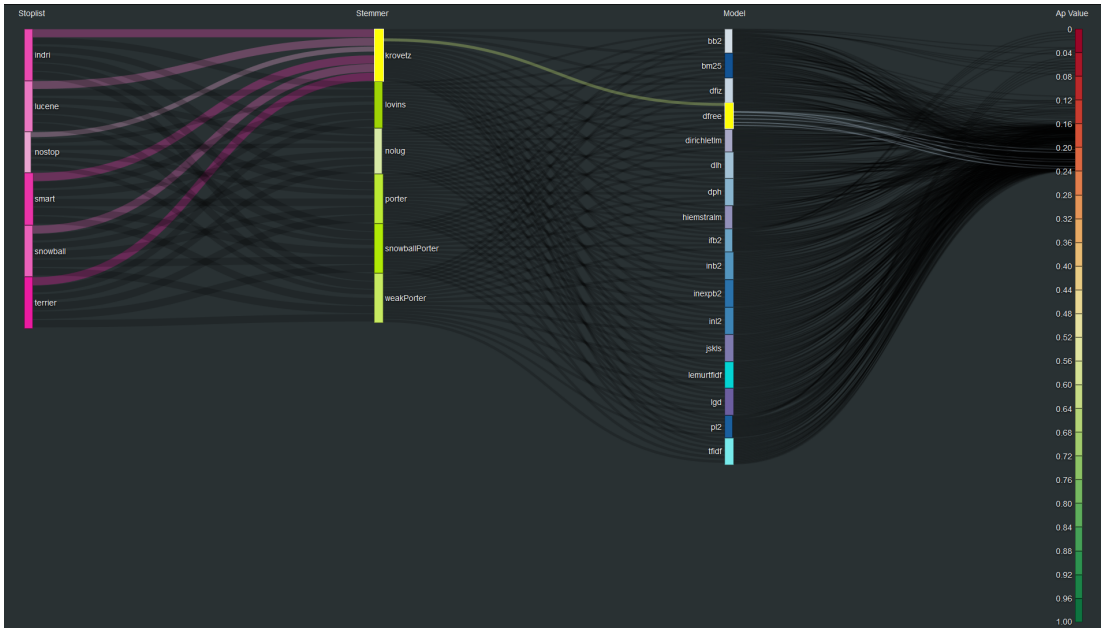


Figura 4.3. *Area di analisi ed esplorazione dei dati di valutazione.*

4.1.1 Selezione dei parametri

Attraverso l'area di selezione dei parametri l'utente può scegliere ciò che vuole realmente visualizzare e come desidera visualizzarlo:

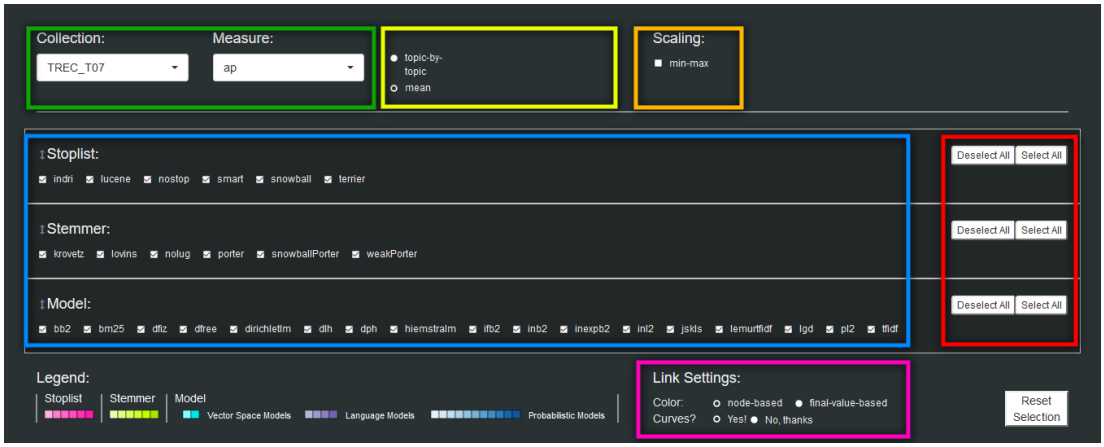


Figura 4.4. *L'area di selezione dei parametri è caratterizzata da sei sezioni principali, qui evidenziate con dei rettangoli.*

1. In alto a sinista (rettangolo verde in figura 4.4) è possibile selezionare una collezione sperimentale (tra Trec7, Trec8, Trec9, Trec10, Trec14 e Trec15) attraverso l'utilizzo di una select list a discesa; allo stesso modo è possibile modificare la metrica utilizzata per la valutazione dei sistemi di IR (figura 4.5a). Le metriche sono: AP, RBP, P10, nDCG, Twist ed ERR.

2. Di default viene considerata un'analisi per l'intera collezione. È possibile visualizzare i dati per singoli topic mediante la selezione del radio button *topic-by-topic* (rettangolo giallo in figura 4.4) che attiva automaticamente la comparsa di una select list a discesa per la scelta del topic da analizzare (figura 4.5b).

3. L'utente può selezionare per ciascuna tipologia di componenti quali visualizzare attraverso l'utilizzo di una serie di checkbox (rettangolo blu in figura 4.4). Sono presenti due pulsanti per la selezione o deselezione totale dei vari componenti per le tre diverse tipologie, ovvero stop list, stemmer e modelli (riquadro rosso in figura 4.4).

(a) *Selezione della visualizzazione dei dati di valutazione per l'intera collezione.*

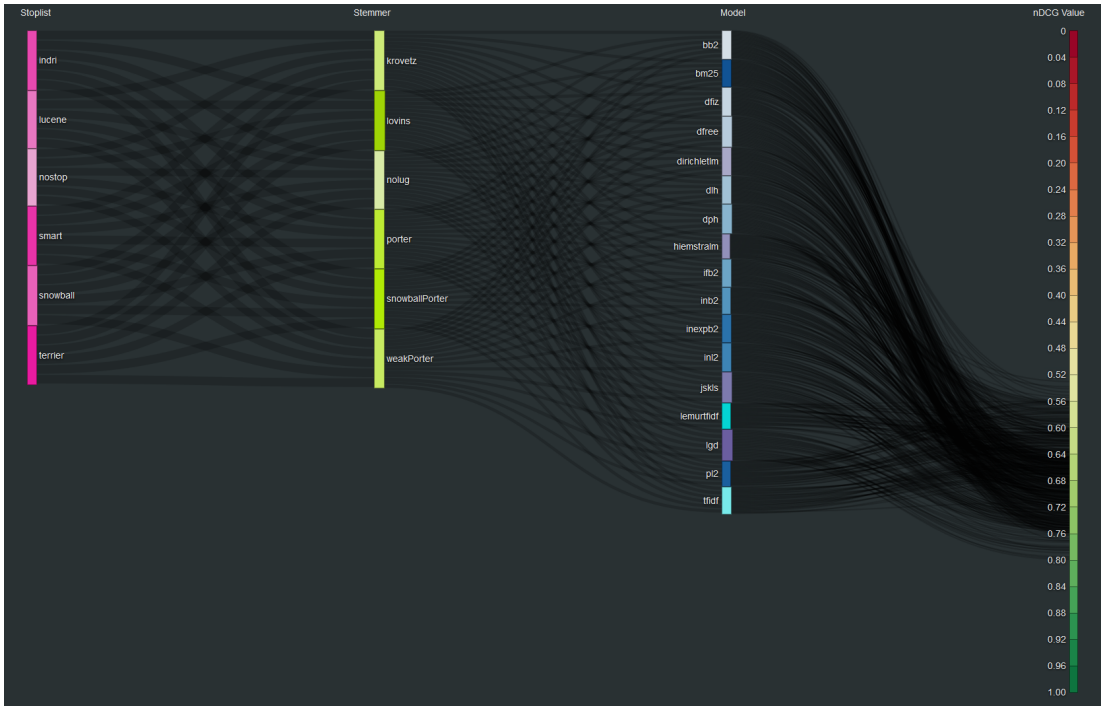
(b) *Selezione della visualizzazione dei dati di valutazione per uno specifico topic.*

Figura 4.5. *Particolare della sezione di selezione dei parametri che evidenzia, oltre alle due select list per la scelta della collezione e della metrica di valutazione, i due radio button per permettere all'utente la scelta di una visualizzazione dei dati di valutazione per l'intera collezione o per uno specifico topic.*

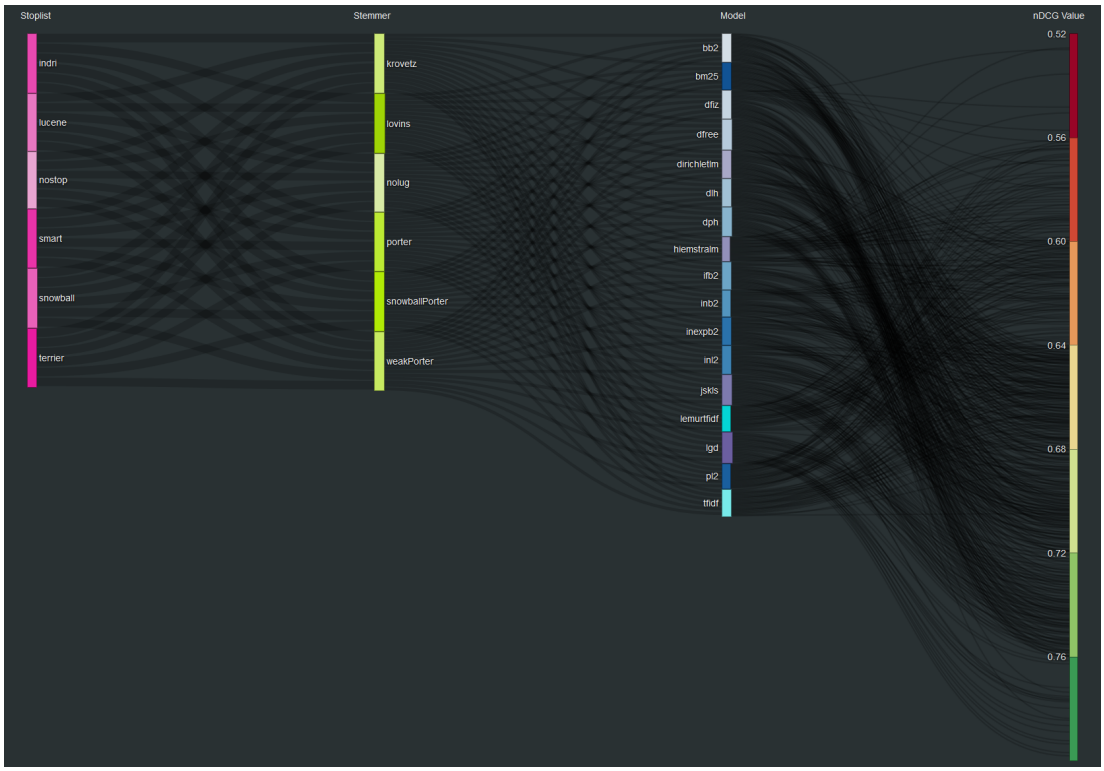
4. È possibile invertire il posizionamento delle famiglie di componenti attraverso una funzione di drag & drop (figura 4.6). Questo è utile in fase di analisi dei dati di valutazione ad esempio per cercare di evidenziare eventuali effetti di interazione tra componenti, difficilmente individuabili altrimenti.

Figura 4.6. *Esempio di riordinamento delle famiglie di componenti. Con questo ordinamento nell'area di visualizzazione saranno rappresentati prima i modelli, poi gli stemmer e infine le stoplist. Di base il posizionamento segue l'ordine stoplist-stemmer-modelli.*

5. Per facilitare l'analisi ed esplorazione dei dati, attraverso la selezione della funzione di scaling *min-max* (riquadro arancione in figura 4.4), si limita la visualizzazione ai soli intervalli di valori dei dati di valutazione compresi tra il minimo e il massimo valore delle performance ottenute per la collezione o lo specifico topic considerato. Questo è utile per permettere una miglior distinzione dei link finali, rappresentati i dati di valutazione dei sistemi di reperimento (figura 4.7).



(a) *Visualizzazione dei dati senza aver applicato la funzione di scaling min-max.*



(b) *Visualizzazione dei dati avendo applicato la funzione di scaling min-max.*

Figura 4.7. *Esempio di visualizzazione dei dati di valutazione per la collezione TREC-T09, topic 455 e metrica di valutazione nDCG. Applicando la funzione di scaling min-max, come evidenziato in figura (b), si riescono a distinguere meglio i link finali. Inoltre c'è una maggior differenziazione di colore tra i vari nodi finali.*

6. Attraverso una coppia di radio button (riquadro rosa in figura 4.4), l'utente può decidere come devono essere colorati i link a seguito della selezione di un nodo. Se viene scelta una colorazione *node-based* i link assumono il colore del nodo d'origine. Questo permette di individuare facilmente il nodo sorgente di un arco evidenziato. Alternativamente può essere scelta una colorazione *final-value-based* che, a seguito di una selezione, assegna un colore al link in base ai valori finali (ovvero i valori di performance) ottenuti dai sistemi che rappresenta.

In questa sezione è anche presente una *legenda* dei colori scelti per ogni singolo componente (figura 4.8), in particolare:

- Le stoplist sono rappresentate utilizzando una scala di colori che va dal rosa tenue ad un rosa vivo. In particolare la tonalità scelta è indicativa circa il numero di stopword presenti all'interno della stop list; un rosa chiaro evidenzia un numero minore di stopword, al contrario un rosa scuro indica una stoplist contenente un numero di stopword elevato.

- I vari stemmer sono caratterizzati da una scala di colore che va da un verde chiaro ad un verde-lime acceso. In verde più chiaro sono rappresentati gli stemmer meno aggressivi, mentre in verde più scuro quelli più aggressivi.

- I modelli sono differenziati in base alla tipologia. Vengono rappresentati: in due diverse tonalità di ciano i due modelli vettoriali, in una scala di viola i language model e per finire in una scala di blu i modelli probabilistici.

I dati visualizzati attraverso l'interfaccia qui di seguito descritta vengono aggiornati dinamicamente ad ogni singolo cambiamento delle impostazioni da parte dell'utente.

Figura 4.8. *Dettaglio della legenda. Si visualizza il nome del componente mediante un tooltip quando il puntatore viene posizionato sopra il quadrato colorato che lo rappresenta. Nell'esempio il quadrato blu scuro rappresenta il modello probabilistico bm25.*

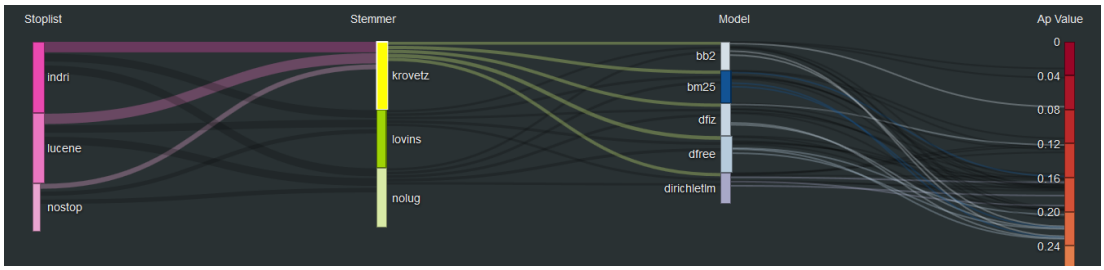
4.1.2 Analisi e valutazione dei sistemi

L'area di visualizzazione dei dati di valutazione dei sistemi è caratterizzata da quattro colonne di valori categorici connessi tra loro attraverso dei link. Le prime tre colonne rappresentano le tre famiglie di componenti e di default sono impostate nell'ordine da sinistra a destra come: stoplist, stemmer e modelli. Ciscun rettangolo rappresenta un singolo componente di un sistema di IR, quindi un link che collega due componenti rappresenta una combinazione o interazione tra questi.

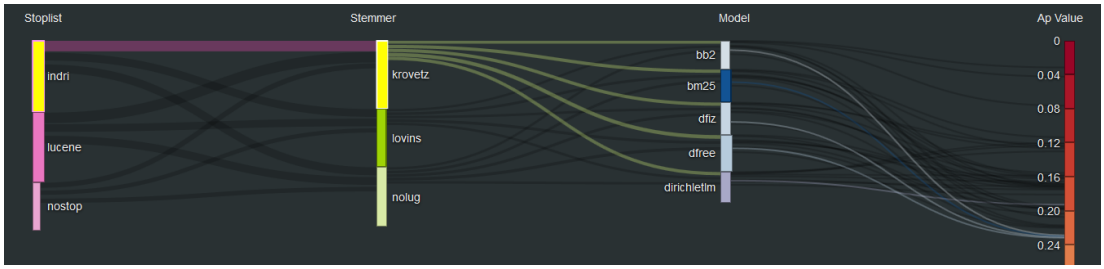
La colonna di destra include di base venticinque rettangoli di ugual dimensione, ciascuno rappresentate un intervallo di valori di ampiezza 0.04 per la metrica di valutazione considerata (tutti gli intervalli sono tra loro disgiunti, scelti in modo che l'unione rappresenti l'intero intervallo unitario $[0, 1]$). Il colore di ciascun rettangolo viene assegnato in base all'intervallo che rappresenta, attraverso uno schema di colori *Rosso-Giallo-Verde*. Un rettangolo che identifica un intervallo vicino al valore zero (basse performance) assumerà una tonalità di rosso, viceversa un rettangolo con intervallo vicino al valore massimo (alte performance) assumerà una tonalità di verde. Ciascun link incidente ad uno di questi insiemi finali rappresenta uno dei 612 sistemi di IR, generato dalla combinazione di una stoplist, uno stemmer ed un modello. Ognuno di questi link è incidente al rettangolo che identifica l'intervallo contenente il valore di performance ottenuto dal sistema che rappresenta, per la metrica di valutazione scelta dall'utente.

Un singolo sistema di reperimento dell'informazione viene rappresentato da un *percorso*, ovvero una successione di link che collegano una stoplist, uno stemmer, un modello ed un valore finale rappresentante la categoria che

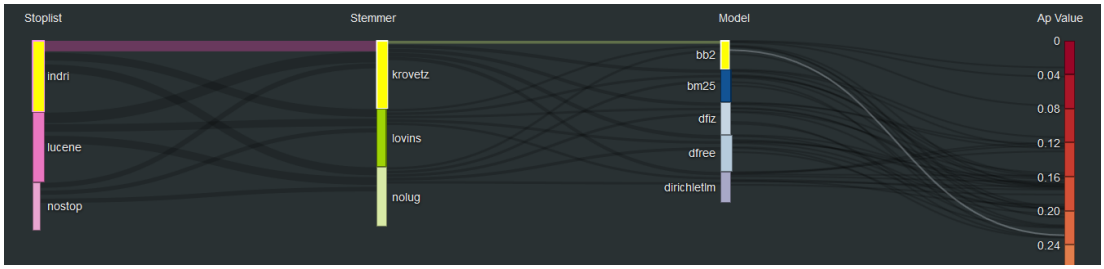
identifica la performance ottenuta dal sistema. L'utente può selezionare un insieme di componenti per evidenziare i percorsi rappresentanti i sistemi ai quali è interessato, come mostrato in figura 4.9.



(a) *In questo esempio è stato selezionato lo stemmer Krovetz. Vengono messi in evidenza tutti i sistemi che utilizzano tale stemmer.*



(b) *In questo esempio sono selezionati lo stemmer krovetz insieme alla stoplist indri, di conseguenza si evidenziano tutti i percorsi rappresentati tutti i sistemi che utilizzano questi due componenti.*



(c) *In questo esempio, oltre allo stemmer krovetz e alla stoplist indri viene selezionato anche il modello bb2. Questo mette in risalto un solo sistema, ovvero il sistema indri-krovetz-bb2, il quale ottiene un valore di AP nell'intervallo* $[0.20, 0.24)$.

Figura 4.9. *Esempio di selezione dei componenti. Si è limitata la visualizzazione a tre stoplist (indri, lucene, nostop), tre stemmer (krovetz, lovins, nolug) e cinque modelli (bb2, bm25, dfiz, dfree, dirichletlm). I dati visualizzati sono relativi alla collezione TREC-T07 e la metrica di valutazione considerata è AP.*

Nell'esempio in figura 4.9a è stato selezionato lo stemmer *krovetz*, perciò vengono evidenziati tutti i percorsi che rappresentano i sistemi che utilizzano tale stemmer (tra quelli visualizzati). In figura 4.9b oltre a *krovetz* è stata

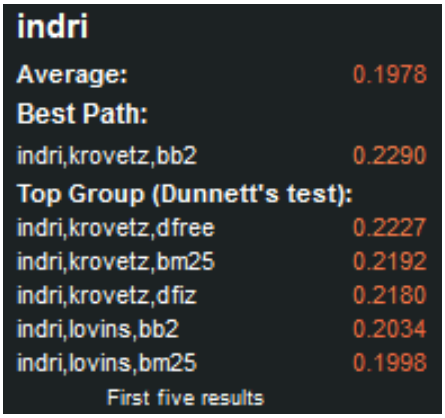
selezionata la stoplist *indri*, quindi si evidenziano tutti i sistemi composti da questi due componenti. Infine in figura 4.9c, poiché è stato selezionato un componente per ciascuna tipologia, si evidenzia un solo sistema, ovvero il sistema *indri-krovetz-bb2* che ottiene un valore di Mean Average Precision compreso nell'intervallo $[0.20, 0.24)$.

Per aiutare l'utente nell'analisi dei dati di valutazione rappresentati, vengono visualizzate tramite dei tooltip alcune informazioni utili, sia quando il puntatore si ferma su un nodo rappresentate un componente, sia quando il puntatore si posiziona su un link. Si evidenziano in particolare tre tipologie di tooltip:

- *Tooltip associato ai nodi:* viene visualizzata una media aritmetica calcolata considerando tutte le performance ottenute dai sistemi di IR che utilizzano quello specifico componente, il sistema che ottiene la miglior performance con il relativo valore di valutazione e una lista dei sistemi (limitata al massimo a cinque) che, secondo il test statistico di Dunnett, non risultano significativamente differenti dal miglior sistema (figura 4.10a).

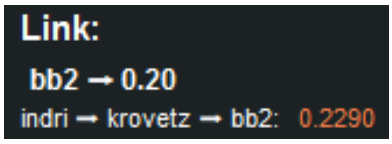
- *Tooltip associato ai link che collegano due componenti:* viene visualizzata una media aritmetica calcolata considerando tutte le performance ottenute dai sistemi di IR che utilizzano i due componenti collegati dal link, il sistema che ottiene la miglior performance e la lista dei sistemi che non risultano significativamente differenti dal migliore (figura 4.10b).

- *Tooltip associato ai link finali:* visualizza l'intero percorso, ovvero il sistema completo, accompagnato dalla performance ottenuta per la metrica considerata (figura 4.10c).



(a) *Tooltip associato al nodo indri.*

(b) *Tooltip associato al link nostop-nolug.*



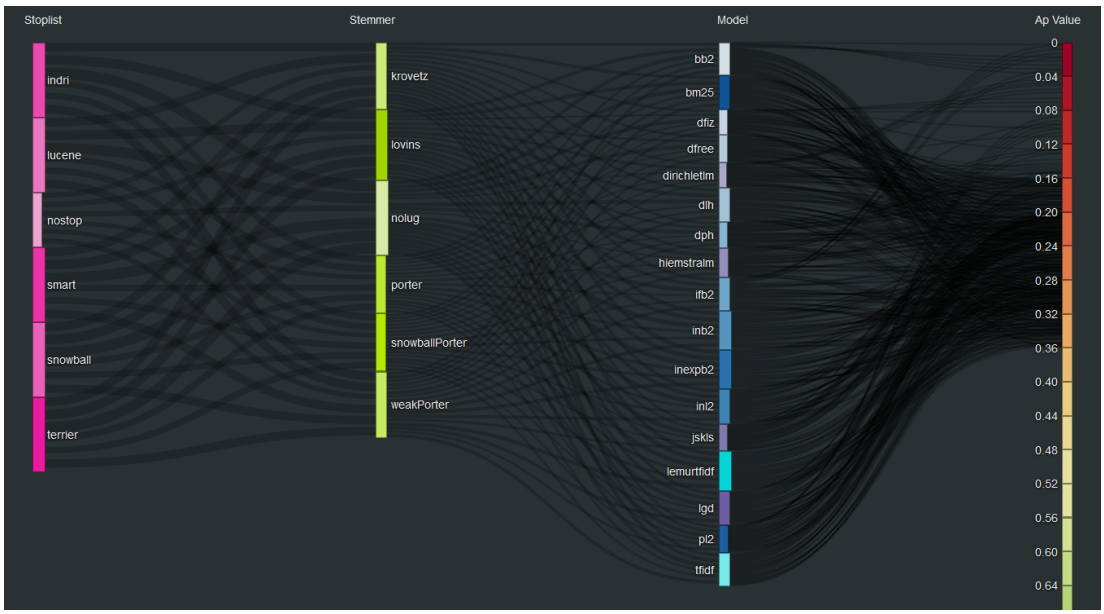
(c) *Tooltip associato al link finale che identifica il sistema indri-krovetz-bb2.*

Figura 4.10. *Esempi di tooltip che possono venire visualizzati dall'utente. Questi esempi sono relativi alla collezione TREC-T07 con AP come metrica selezionata.*

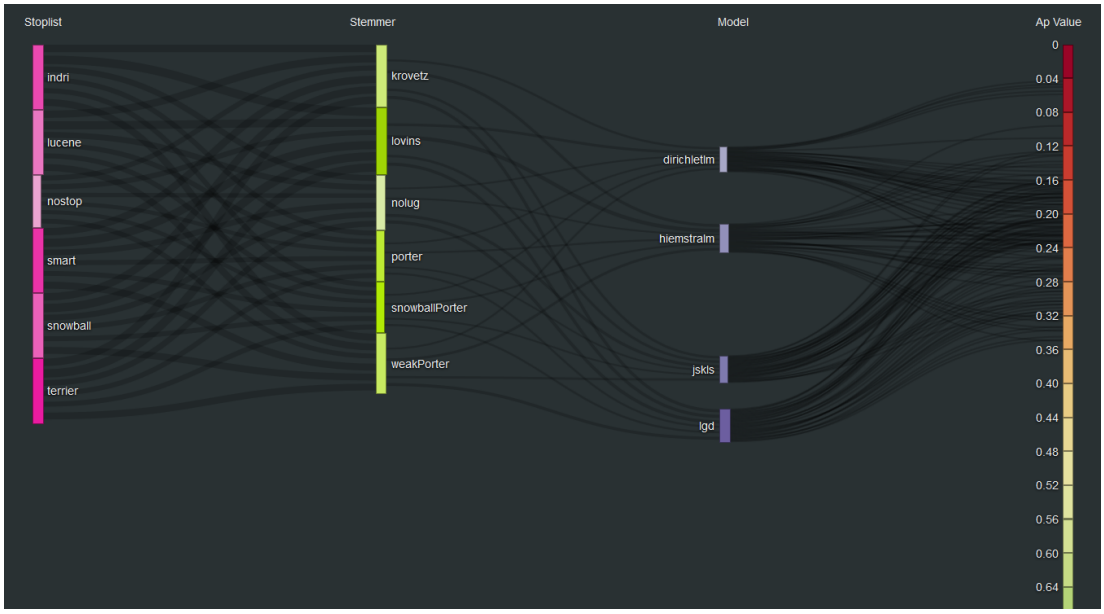
4.1.3 Tecniche di visual analytics

SANKEY utilizza tecniche di visual analytics per supportare l'utente nell'esplorazione dei dati e per aiutarlo alla comprensione di alcune caratteristiche particolari dei sistemi di IR visualizzati.

Per ciascun nodo viene calcolata una media aritmetica dei valori di performance ottenuti dai sistemi di IR che utilizzano quello specifico componente. Per permettere all'utente di determinare in modo immediato quale componente (tra quelli appartenenti alla stessa famiglia) garantisce i migliori risultati per la specifica collezione di documenti presa in considerazione, la dimensione del nodo dà una indicazione circa il valore della media. Se il nodo è di dimensione maggiore rispetto agli altri nodi della stessa famiglia allora il componente associato ha un effetto positivo, ovvero la media sarà maggiore rispetto a quelle degli altri componenti. Un nodo di dimensione più piccola avrà associata una media inferiore a quella degli altri componenti appartenenti allo stesso gruppo.



(a) *Visualizzazione di tutti i sistemi per il topic 353 della collezione TREC-T07, considerando AP come metrica di valutazione. La dimensione dei rettangoli ci evidenzia come alcuni modelli siano migliori di altri (ad esempio il modello lemurtfidf). Inoltre è possibile notare come nolug abbia un effetto positivo sulle performance rispetto agli altri stemmer, mentre nostop ha un effetto negativo.*



(b) *Visualizzazione dei soli sistemi che utilizzano un language model per il topic 353 della collezione TREC-T07, considerando AP come metrica di valutazione. Rispetto all'esempio precedente le medie sono state aggiornate considerando i soli sistemi visualizzati. In questo caso, osservando le nuove dimensioni dei nodi, si nota che nolug non è la tecnica di stemming preferibile per i soli language model, ma sono diventate krovetz e lovins.*

Figura 4.11. *Esempi di ridimensionamento dei nodi in base ai sistemi visualizzati.*

Allo stesso modo viene associata una media aritmetica ad ogni link che collega due componenti. In questo caso la media viene calcolata tra i valori delle performance ottenute da tutti i sistemi che utilizzano i due componenti connessi dall'arco. La dimensione in larghezza dell'arco è proporzionale alla sua media, quindi una media più alta può identificare una interazione positiva tra i due componenti connessi e ciò viene rappresentato con un link di larghezza maggiore, al contrario l'arco avrà larghezza minore se la media associata sarà piccola rispetto a quelle degli altri archi.

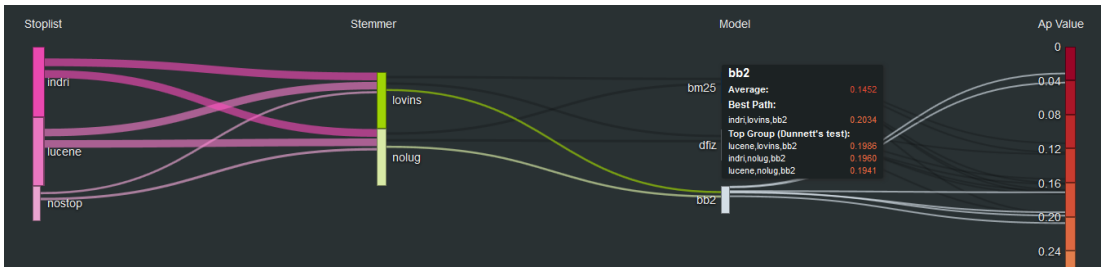
Le medie di nodi e archi si aggiornano dinamicamente ogni qualvolta l'utente decide di visualizzare un numero maggiore o minore di sistemi, andando a selezionare i vari componenti nella sezione di selezione dei parametri. Di conseguenza vengono aggiornate anche le dimensioni di nodi e archi per evidenziare i cambiamenti determinati dall'aggiunta o la rimozione dei sistemi di IR considerati. In figura 4.11 vengono visualizzati i dati di valutazione dei sistemi per la collezione TREC_T07 relative al topic 353, in particolare utilizzando la metrica di valutazione AP. In figura 4.11a, considerando tutti i sistemi, risulta evidente che alcuni modelli garantiscono in media migliori risultati di valutazione rispetto ad altri, poiché i nodi associati hanno una dimensione maggiore sia in altezza che in larghezza. Allo stesso modo *nolug* risulta essere la tecnica di stemming in generale preferibile (media = $0,2844$), mentre l'assenza di una stoplist ha un effetto negativo nelle performance (media = $0,1907$). In figura 4.11b si è limitata l'analisi dei soli sistemi che utilizzano un language model. Per questo insieme di sistemi si evidenzia come la tecnica di stemming preferibile non sia più *nolug* (media = $0,2090$), ma risulti essere *lovins* (media = $0,2548$) oppure *krovetz* (media = $0,2370$).

Nel caso l'utente sia interessato all'esplorazione dei dati di valutazione per un'intera collezione, SANKEY applica un test di confronto multiplo per determinare in un insieme di dati di valutazione di sistemi di IR, quali differiscono significativamente rispetto al sistema, tra quelli considerati, che ottiene la miglior performance. Il test viene applicato:

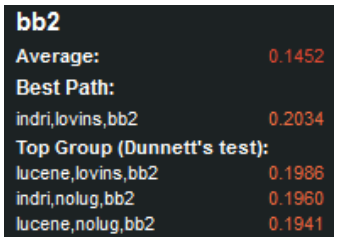
- quando si analizza un singolo componente (posizionando il cursore sul nodo che la rappresenta), considerando le performance ottenute dai sistemi che utilizzano tale componente;

- quando si analizza una coppia di componenti (posizionando il cursore sul link che li unisce), considerando le performance ottenute da tutti i sistemi che utilizzano i due componenti presi in esame;

Per fare ciò viene utilizzato il test statistico di Dunnett, proposto per permettere il confronto di due o più trattamenti con un trattamento di controllo, al fine di determinare se esiste una differenza significativa tra la media di controllo e tutte le altre considerate [Dunnett, 1955].



(a) *Esempio che evidenzia il funzionamento del test di Dunnett.*



(b) *Dettaglio del tooltip.*

Figura 4.12. *Esempio che evidenzia come il test di Dunnett escluda dal top group alcuni sistemi, ovvero il sistema nostop-lovins-bb2 e il sistema nostop-nolug-bb2 che ottengono performance rispettivamente di* $0,03$ *e* $0,0491$ *(selezionando la metrica di valutazione AP). Queste sono considerate significativamente differenti dalla performance ottenuta dal sistema indri-lovins-bb2, pari a* $0,2034$.

Dato il miglior sistema tra quelli considerati, ovvero quello con la miglior performance, il test di Dunnett viene utilizzato per determinare quali dei rimanenti sistemi ottengono performance significativamente minori, applicando un livello $\alpha = 0,05$ (dove α rappresenta la probabilità di effettuare

un errore statistico di I specie). I sistemi appartenenti a questo insieme non vengono elencati nel tooltip mostrato all'utente, poiché non sono considerati appartenenti al *top group*, ovvero al gruppo di sistemi migliori. Un esempio di utilizzo del test di Dunnett viene mostrato in figura 4.12.

4.1.4 Esempi di utilizzo del sistema SANKEY

Vengono ora forniti alcuni esempi di utilizzo del sistema SANKEY per poter descrivere al meglio alcune sue funzionalità.

- **Task d'esempio:** Determinare per la collezione TREC-T15 qual è il miglior modello, considerando la metrica di valutazione Twist.

Per la risoluzione del task è necessario selezionare la collezione TREC-T15 e la metrica di valutazione Twist. La collezione viene selezionata attraverso la select list a discesa denominata "*Collection*". Durante la selezione vengono date alcune informazioni sulla collezione che si sta scegliendo. Un tooltip mostra il nome della collezione, l'anno di creazione, il tipo di collezione (news search o web search) ed altri dettagli riguardanti il corpus, i topic e il ground truth. Attraverso la select list a discesa denominata "*Measure*" è possibile selezionare la metrica Twist (figura 4.13).

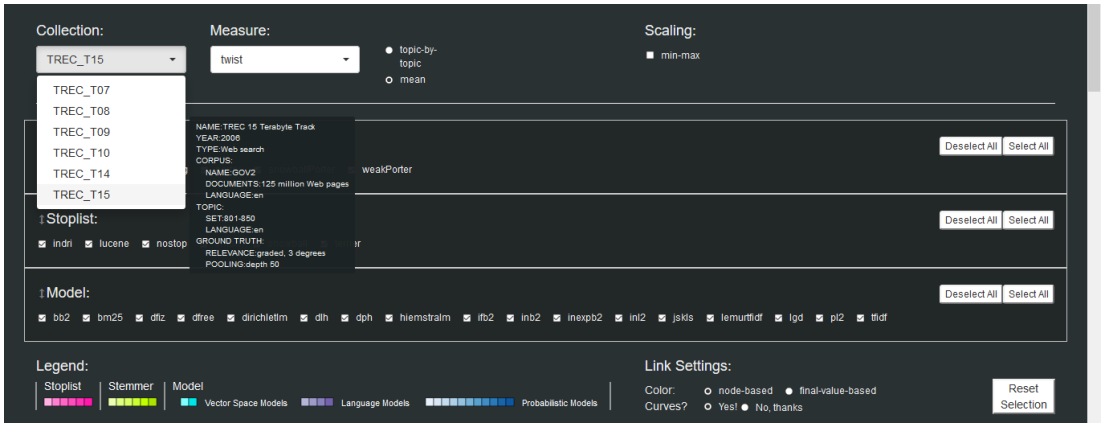


Figura 4.13. *Selezione della collezione TREC-T15 e della metrica Twist. Durante la selezione della collezione sperimentale è possibile visualizzare alcune informazioni relative alla collezione grazie ad un tooltip.*

A questo punto la fase di selezione dei parametri può definirsi conclusa, e si può passare all'analisi dati visualizzati. Per identificare il miglior modello è possibile analizzare le medie e quindi le dimensioni dei nodi rappresentati i diciassette modelli (figura 4.14).

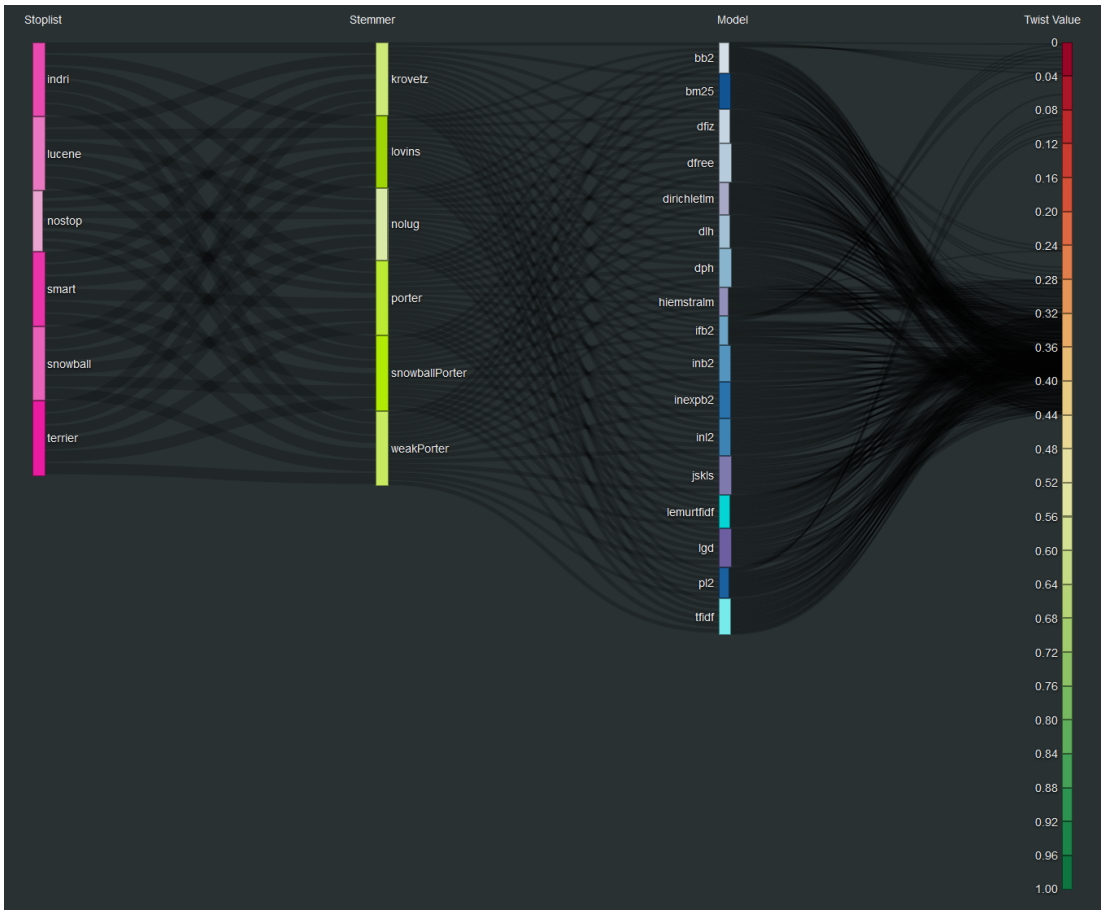


Figura 4.14. *Visualizzazione dei dati di valutazione per la collezione TREC-T15 e metrica Twist. Dalle differenti dimensioni dei nodi è possibile individuare quali sono i modelli che ottengono le migliori performance. I modelli migliori sono dfree, dph, jskls ed lgd.*

Dalla dimensione risulta che quattro modelli ottengono in media performance superiori agli altri, tali modelli sono dfree, dph, jskls ed lgd. Per determinare quali di questi quattro modelli è effettivamente il migliore, è sufficiente analizzare i tooltip per ciascuno di essi (figura 4.15). Attraverso i tooltip è possibile identificare le medie:

- dfree: media = *0,4074*;

- dph: media = *0,4094*;
- jskls: media = *0,4130*;
- lgd: media = *0,3999*.

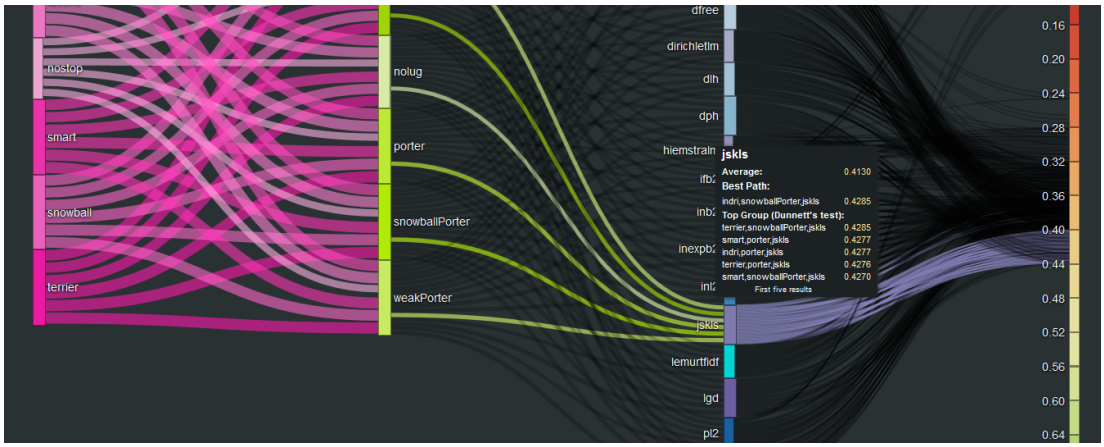


Figura 4.15. *Passando con il puntatore sopra ad un nodo è possibile visualizzare alcune informazioni attraverso un tooltip. Nell'immagine si evidenzia il tooltip associato al modello jskls.*

Il modello jskls è il miglior modello, quello che ottiene le migliori performance indipendentemente dalla stoplist o lo stemmer scelto. Scegliendo di visualizzare solo il modello jskls (utilizzando il pulsante *deselect all* per la lista di modelli e selezionando la checkbox associata a jskls) possiamo ricavare altre interessanti informazioni (figura 4.16).

Figura 4.16. *Nell'esempio si vuole limitare la visualizzazione del solo modello jskls. È conveniente deselezionare tutti i modelli tramite il pulsante "deselect all" sulla destra e solo successivamente selezionare il modello jskls tramite la relativa checkbox.*

Le dimensioni dei nodi rappresentanti i vari stemmer e le varie stoplist risultano molto simili tra loro ed i valori di performance dei sistemi visualizzati sono quasi tutti compresi nell'intervallo $[0.40, 0.44)$ ad eccezione dei sistemi nolug-snowball-jskls (twist = *0,3998*) e nolug-lucene-jskls (twist = *0,3986*), che ottengono valori molto vicini a tale intervallo (figura 4.17). Di conseguenza

è possibile concludere che il modello jskls risente poco di variazioni di performance dovute alla scelta di un particolare stemmer, di una particolare stoplist o di una particolare combinazione di questi due componenti. Portando il puntatore del mouse sul nodo che rappresenta il modello, è possibile visualizzare il migliori sistemi in assoluto che corrispondono a indri-snowballPorter-jskls e terrier-snowballPorter-jskls (twist = $0,4285$).

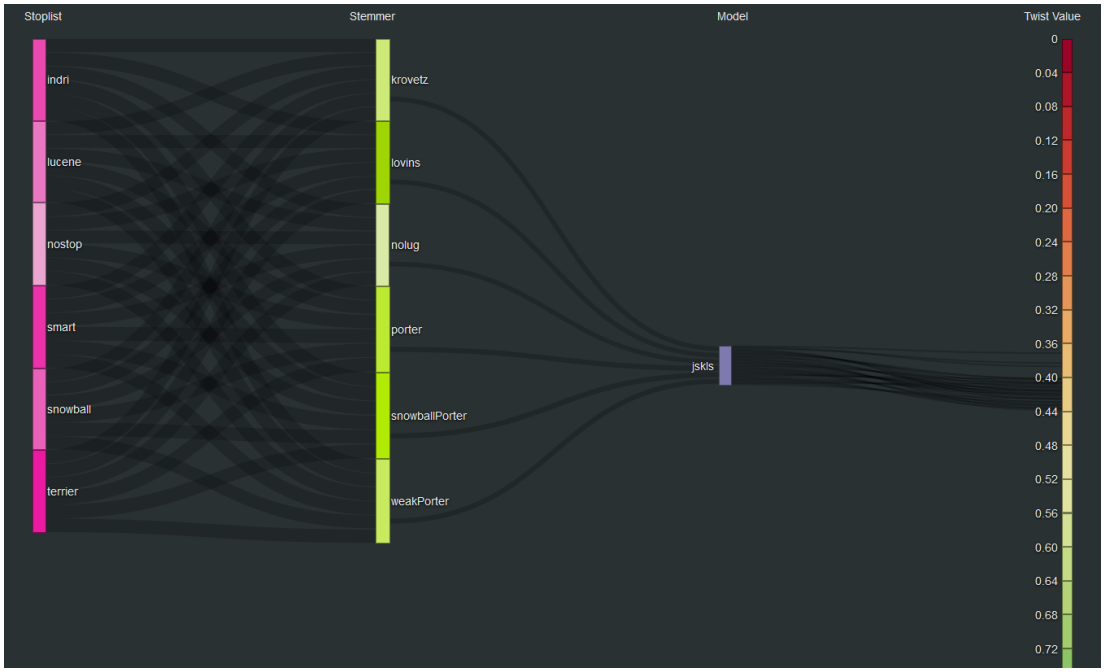


Figura 4.17. *Nell'immagine vengono visualizzati i sistemi che utilizzano il modello jskls. Le dimensioni dei nodi rappresentanti i vari stemmer sono molto simili tra loro e lo stesso si può dire per le varie stoplist. Inoltre, poiché i valori di performance dei sistemi visualizzati sono quasi tutti compresi nell'intervallo* $[0.40, 0.44)$*, si può concludere che le performance ottenute dal modello jskls non risentono dell'utilizzo di una particolare stoplist e/o di un particolare stemmer.*

- **Task d'esempio:** Per la collezione TREC-T08, utilizzando la metrica di valutazione AP, determinare in che condizioni si ottengono basse performance.

Per prima cosa è necessario selezionare la collezione TREC-T08 e la metrica di valutazione AP, come fatto in precedenza. Osservando i dati visualizzati (figura 4.18) è possibile notare come gran parte dei sistemi ottengano

performance comprese negli intervalli$[0.16, 0.20)$, $[0.20, 0.24)$ o $[0.24, 0.28)$, poiché molti dei link finali puntano ai tre nodi rappresentanti questi tre intervalli.

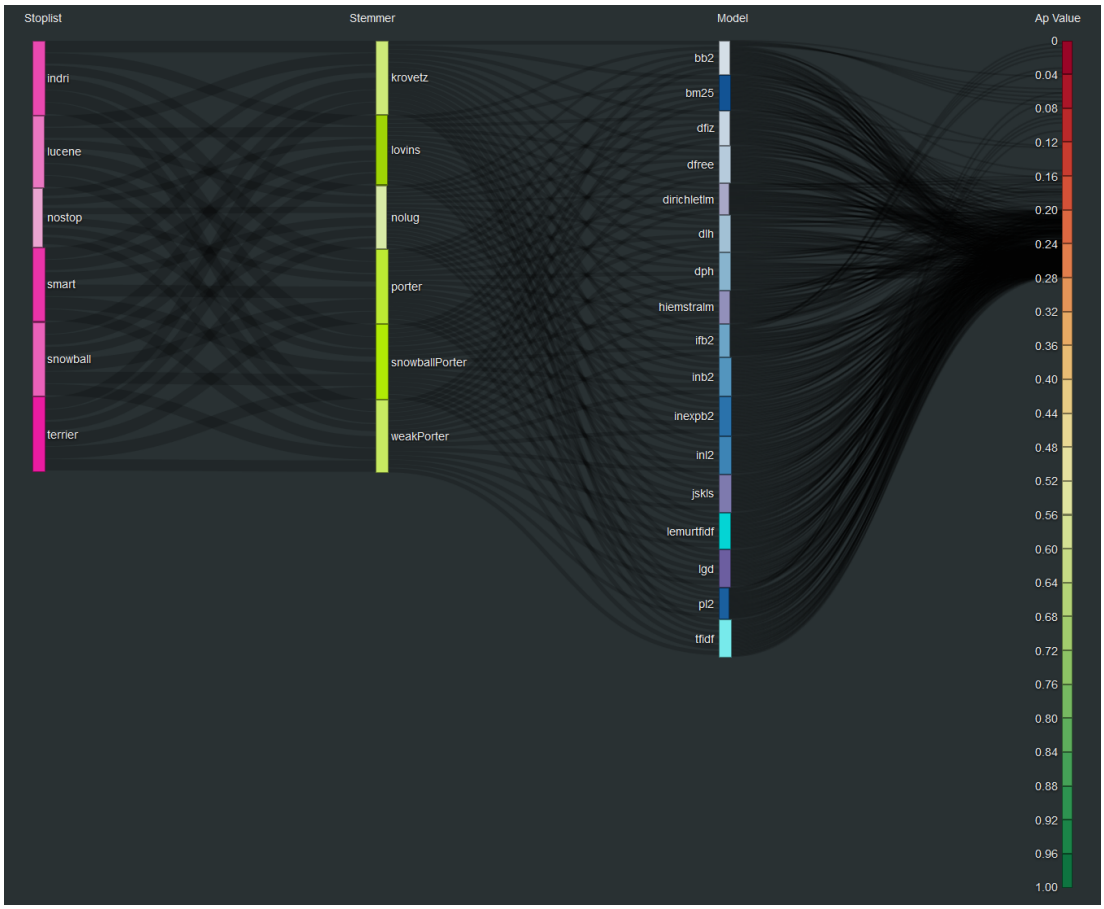


Figura 4.18. *Visualizzazione dei dati di valutazione per la collezione TREC-T08 e metrica AP. Osservando i link finali si nota che gran parte dei sistemi ottengono performance comprese negli intervalli* $[0.16, 0.20)$, $[0.20, 0.24)$ *o* $[0.24, 0.28)$.

Tutti i sistemi che ottengono un valore di AP inferiore a $0,16$ possono essere considerati sistemi a "basse performance". È possibile evidenziare questi sistemi selezionando i nodi finali come mostrato in figura 4.19. I percorsi evidenziati ci permettono di comprendere che soltanto alcuni modelli ottengono performance inferiori a $0,16$ e sono i modelli bb2, bm25, dfiz, ifb2 e pl2.

Osservando con attenzione si riescono a ricavare altre informazioni dai percorsi evidenziati, ma per faciliate meglio l'esplorazione dei dati è

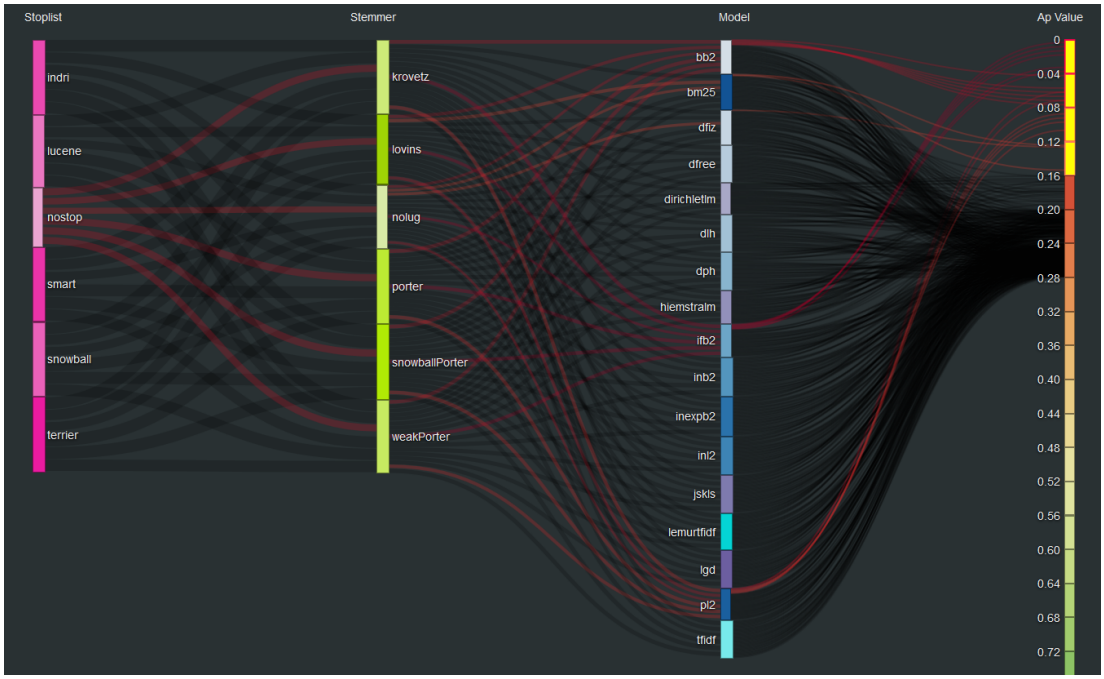


Figura 4.19. *Vengono selezionati quattro intervalli finali per evidenziare tutti i sistemi che ottengono performance inferiori al valore* $0,16$.

conveniente spostare l'ordinamento dei componenti attraverso la funzione di drag & drop sui riquadri che permettono la selezione dei componenti per ciascuna famiglia. Ad esempio è possibile spostare le stoplist in fondo alla lista, in modo da evidenziare meglio se le basse performance dipendono in particolare da una specifica stoplist. Osservando il nuovo grafico ottenuto (figura 4.21) è facile verificare come le basse performance siano dovute alla mancanza di utilizzo di una stoplist (nostop). Spostando invece gli stemmer sulla destra (esattamente come è stato fatto in precedenza per le stoplist) si nota come i sistemi evidenziati non ottengano basse performance a causa dell'utilizzo di un particolare stemmer, infatti tutti gli stemmer vedono dei percorsi evidenziati.

Quindi per la collezione TREC-T08 e metrica AP si ottengono basse performance quando non viene utilizzata nessuna stoplist con i modelli bb2, bm25, dfiz, ifb2 e pl2. I colori dei link evidenziano bb2 e ifb2 come i modelli che più risentono della mancanza di stoplist, essendo quelli che ottenengono valori di AP più bassi (link di colore rosso).

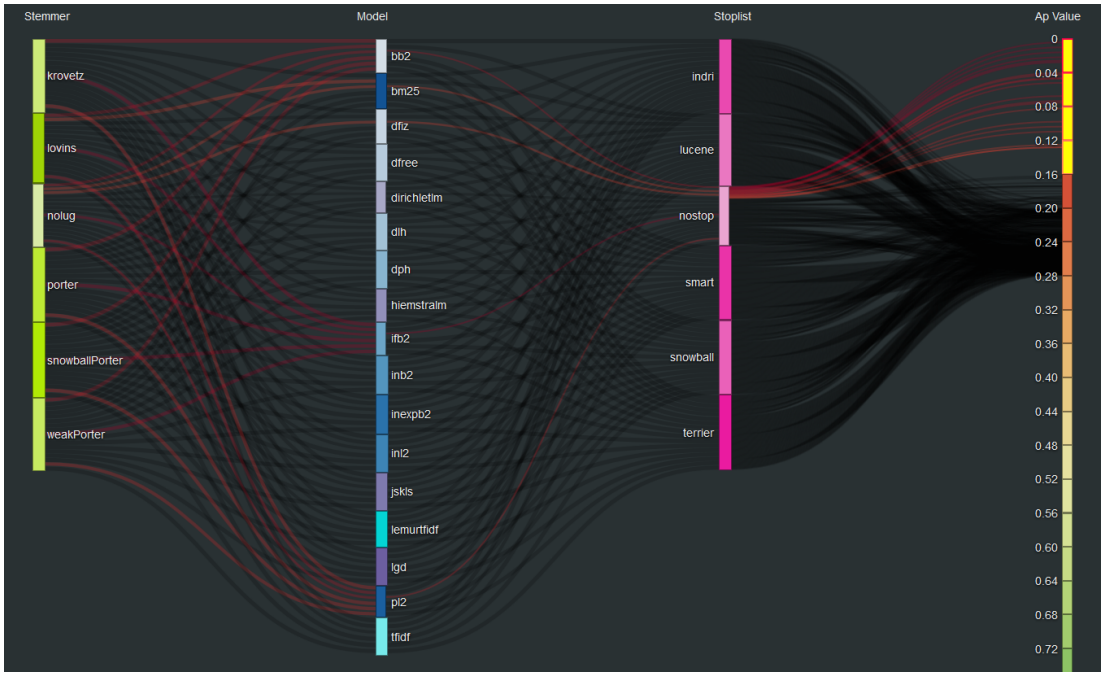


Figura 4.20. *Visualizzazione dei dati di valutazione dopo lo spostamento delle stoplist in ultima posizione. In questo modo si riesce a capire che basse performance sono ottenute a causa della mancanza di una stoplist (nostop).*

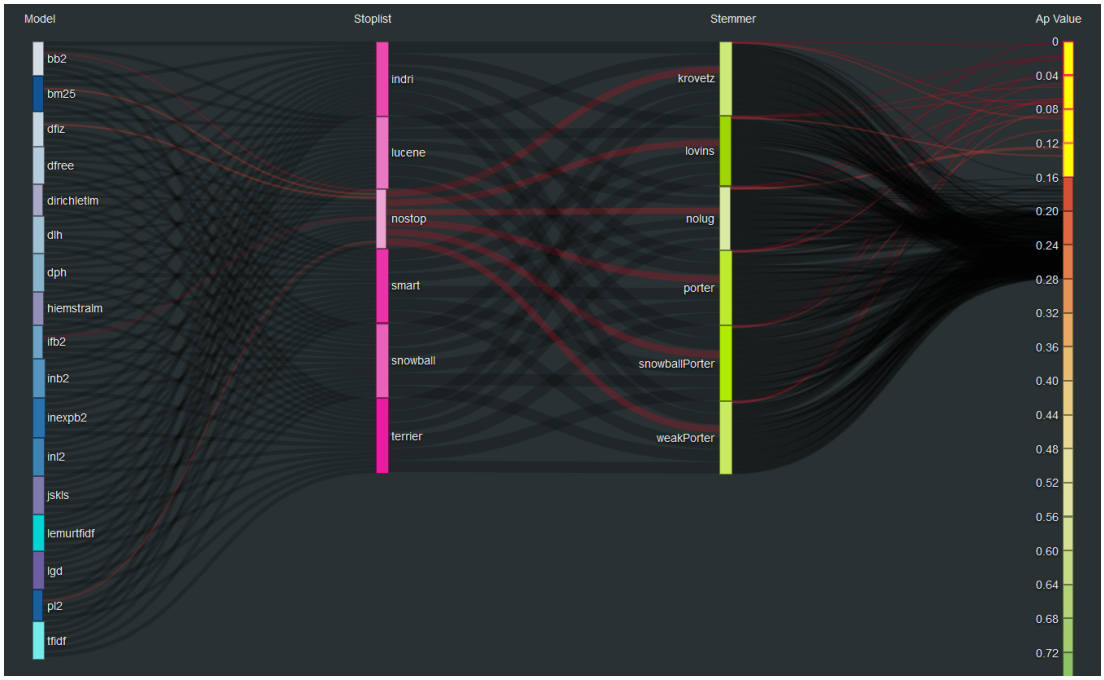


Figura 4.21. *Visualizzazione dei dati di valutazione dopo lo spostamento delle stemmer in ultima posizione. Non è la scelta di qualche stemmer particolare a causare basse performance.*

5 Validazione

Il metodo principale per valutare uno strumento di IV ed evidenziarne eventuali criticità è attraverso un *test d'utente* o *test di usabilità*. Un test di usabilità viene svolto da una serie di utenti neutrali e imparziali, invitati a testare lo strumento sviluppato. Lo scopo del test consiste nell'analisi del comportamento dell'utente per evidenziare difficoltà riscontrate durante lo svolgimento di alcuni compiti o *task* proposti e pensati appositamente per spingerlo ad utilizzare il sistema nella sua interezza. Idealmente, è preferibile effettuare un test di usabilità con un elevato numero di utenti per poter ricavare statistiche più accurate e per riuscire a scoprire il maggior numero di problemi del sistema testato. Tuttavia è stato dimostrato che cinque utenti sono sufficienti per evidenziare circa l'80% dei problemi di una interfaccia [Nielsen and Landauer, 1993]. Al test di usabilità del sistema SANKEY hanno partecipato nove studenti che in passato hanno frequentato il corso di reperimento dell'informazione e che quindi possiedono le basi fondamentali per comprendere l'obiettivo che ha condotto alla costruzione di strumenti visuali per la valutazione dei sistemi di reperimento dell'informazione e per riuscire a muoversi al meglio nell'interfaccia in quanto pensata per utenti esperti della materia.

5.1 Metodologia

Il test è stato svolto per valutare il sistema SANKEY e per confrontarlo con CLAIRE, un altro strumento visuale per la valutazione dei sistemi di IR [Angelini et al., 2017]. Si è suddiviso il test in tre fasi principali:

1. una fase di spiegazione delle funzionalità dei due strumenti visuali che gli utenti avrebbero dovuto utilizzare per risolvere alcuni task proposti;
2. una fase di risoluzione di tre task utilizzando prima uno dei due sistemi e poi il successivo. A questa fase è seguita la compilazione di due questionari: uno di valutazione dei singoli sistemi (tre domande per ciascun sistema, uguali tra loro) e uno di preferenza (quattro domande);
3. una fase di risoluzione di altri cinque task per il solo sistema SANKEY, al fine di valutare il sistema in ogni aspetto e di evidenziarne il maggior numero di eventuali problemi possibile;

Poiché parte del test consisteva nel confronto di due sistemi, il test è stato suddiviso in due sessioni separate. Nella prima sessione hanno partecipato quattro dei nove utenti i quali, a seguito di una iniziale spiegazione del sistema SANKEY e in seguito del sistema CLAIRE, hanno dovuto risolvere i tre task proposti utilizzando i due sistemi nel seguente ordine: prima SANKEY, poi CLAIRE. Durante la seconda sessione sono stati presentati i due strumenti nell'ordine inverso, prima CLAIRE e successivamente SANKEY. I cinque utenti hanno in seguito risolto i tre task utilizzando prima CLAIRE e poi SANKEY. Questo tipo di divisione si è resa necessaria per evitare di favorire uno dei due sistemi rispetto all'altro. Sebbene presente una fase di spiegazione iniziale per permettere ai partecipanti al test di comprendere al meglio tutte le funzionalità dei due sistemi, questi strumenti di visualizzazione sono caratterizzati da una elevata difficoltà di utilizzo e comprensione poiché integrano numerose funzioni e aspetti avanzati come il test statistico di Dunnett. L'utente prende confidenza con un'interfaccia man mano che la utilizza e la esplora. Poiché le due interfacce rappresentano, anche se in modi

differenti, gli stessi dati, ci si aspettava che l'utente acquisisse una maggior dimestichezza nell'utilizzo del secondo sistema da lui utilizzato, in quanto aveva già avuto modo di risolvere i task con l'altro strumento visuale.

Q1	Quanto ritieni intuitivo l'utilizzo di SANKEY/CLAIRE?
Q2	Quanto ritieni che il sistema visuale SANKEY/CLAIRE possa dare benefici alla comprensione delle prestazioni dei sistemi di IR?
Q3	Quanto il sistema visuale SANKEY/CLAIRE si è rivelato efficace per la risoluzione dei task proposti?

Tabella 5.1. *Questionario di valutazione. Le tre domande sono state poste per la valutazione sia del sistema SANKEY che del sistema CLAIRE.*

A seguito della risoluzione dei primi tre task sono stati proposti due questionari. Il primo, le cui domande sono presentate nella tabella 5.1, è stato pensato per consentire agli utenti di valutare i singoli sistemi visuali. Per ciascuna domanda l'utente ha dovuto indicare un valore compreso nell'intervallo $[1, 5]$ dove ciascun punteggio identificava una specifica risposta: $\{$ 1 = Per nulla, 2 = Poco, 3 = Abbastanza, 4 = Molto e 5 = Estremamente$\}$.

Q1	Quale sistema rappresenta in modo più efficace i dati?
Q2	Quale sistema offre l'interfaccia più intuitiva per interagire con i dati visualizzati?
Q3	Quale sistema si è rivelato più completo per la risoluzione dei task assegnati?
Q4	Quale sistema visuale complessivamente hai preferito utilizzare?

Tabella 5.2. *Questionario di preferenza.*

Il secondo questionario aveva lo scopo di permettere all'utente di assegnare una preferenza tra i due sistemi per quattro specifiche domande. Le quattro domande, presentate nella tabella 5.2, dovevano essere risposte attraverso la scelta di uno dei cinque valori di preferenza. Un esempio viene rappresentato in figura 5.1 dove il valore 0 (zero) rappresenta un giudizio di indifferenza tra i due sistemi, valori a sinistra dello zero stanno ad indicare un indice di preferenza a favore dello strumento visuale SANKEY, mentre i valori a destra dello zero evidenziano indici di preferenza per il sistema CLAIRE,

in particolare i simboli sono etichettati come: { + = Preferibile, ++ = Molto preferibile }.

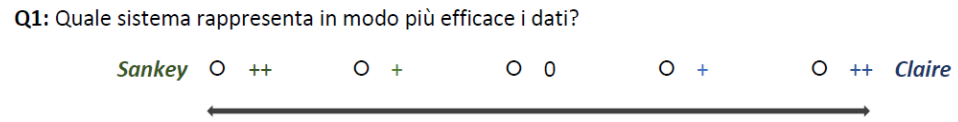


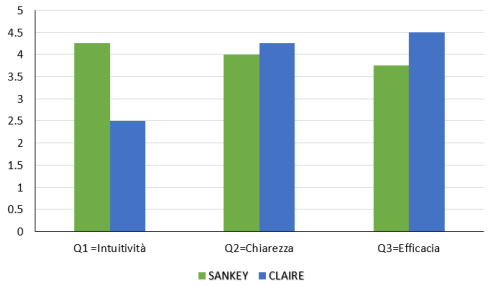
Figura 5.1. *Esempio di domanda del questionario di preferenza che evidenzia le cinque possibili risposte. Il valore* 0 *(zero) rappresenta un giudizio di indifferenza tra i due sistemi, valori a sinistra dello zero stanno ad indicare un indice di preferenza a favore dello strumento visuale SANKEY, mentre i valori a destra dello zero evidenziano indici di preferenza per il sistema CLAIRE. In particolare, il significato assunto dai due simboli è:* { + = *Preferibile*, ++ = *Molto preferibile* }.

Ciascuna di queste domande è stata posta con l'obiettivo di determinare quale dei due sistemi fosse giudicato dagli utenti *più efficace* (Q1), *più intuitivo* (Q2), *più completo* (Q3) e quale dei due sia in generale stato preferito in fase di utilizzo (Q4).

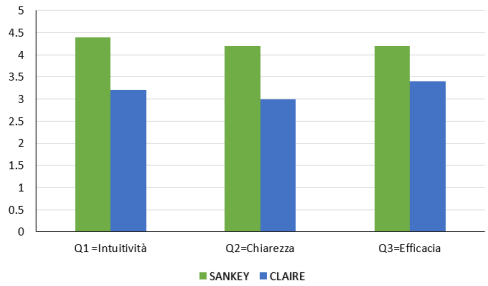
Dopo la compilazione dei due questionari è stato richiesto agli utenti di completare altri cinque task specifici per il sistema SANKEY. In questo caso gli utenti sono stati osservati con attenzione per evidenziare i problemi dell'interfaccia e le difficoltà riscontrate nella risoluzione dei task.

5.2 Risultati

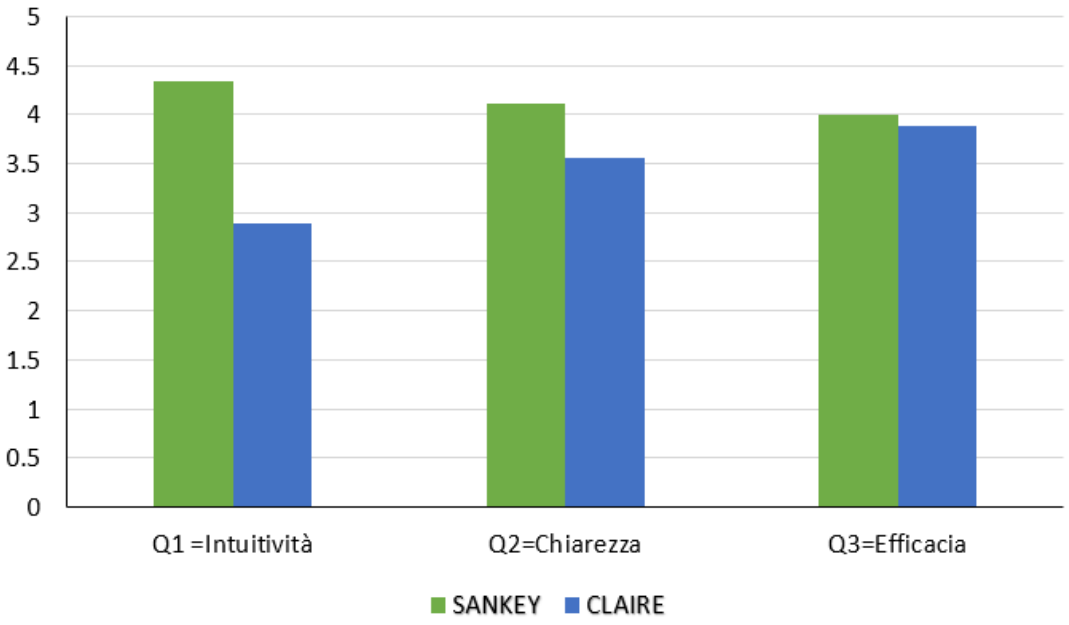
I risultati del primo questionario sono presentati in figura 5.2, dove sono rappresentate le medie ottenute dai due sistemi per ciascuna delle tre domande. Sono presenti anche i grafici specifici di ciascuna sessione per evidenziare eventuali differenze di valutazione dovute all'ordine con la quale i due sistemi sono stati utilizzati. In generale il sistema SANKEY ha ottenuto valutazioni superiori a CLAIRE, risultando molto intuitivo (ottenendo per Q1 media = $4,33$ e deviazione standard = $0,71$), efficace (Q3 con media = $4,0$ e deviazione standard = $0,87$) e in grado di dare numerosi benefici per l'esplorazione dei dati di valutazione dei sistemi di IR (Q2 con media = $4,11$, deviazione standard = $0,33$).



(a) *Istogramma che riporta le medie dei risultati ottenuti (per il primo questionario) nella prima sessione del test di usabilità al quale hanno partecipato quattro utenti. È stato richiesto agli utenti di completare i task utilizzando prima il sistema SANKEY e successivamente CLAIRE.*



(b) *Istogramma che riporta le medie dei risultati ottenuti (per il primo questionario) nella seconda sessione del test di usabilità al quale hanno partecipato cinque utenti. È stato richiesto agli utenti di completare i task utilizzando prima il sistema CLAIRE e successivamente SANKEY.*



(c) *Istogramma che riporta le medie dei risultati ottenuti complessivamente per il primo questionario, tenendo in considerazione le risposte date da tutti e nove gli utenti partecipanti al test di usabilità.*

Figura 5.2. *Confronto dei due sistemi basato sul questionario di valutazione singola.*

I risultati della prima sessione (figura 5.2a), dove è stato utilizzato prima il sistema SANKEY e solo successivamente il sistema CLAIRE, hanno evidenziato una preferenza per CLAIRE, soprattutto per quanto riguarda l'efficacia di rappresentazione, dove CLAIRE ha ottenuto media = $4,5$, contro SANKEY che invece ha ottenuto media = $3,75$. Nella seconda sessione (figura 5.2b), invece, SANKEY ha ottenuto valutazioni superiori in tutte e tre le domande. Ciò evidenzia che, al fine di riuscire ad interagire al meglio con un sistema visuale, è necessaria una lunga fase di allenamento sull'utilizzo del sistema.

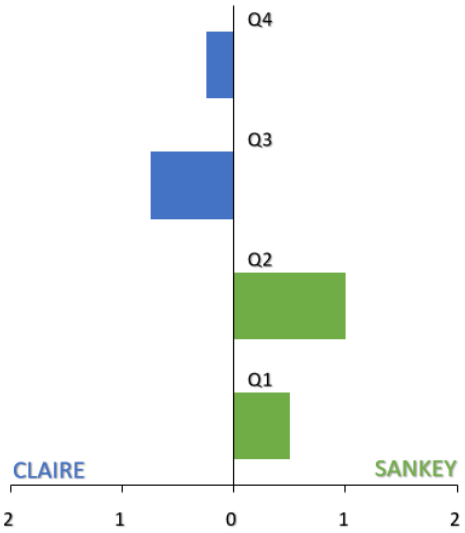
I risultati del questionario di preferenza sono rappresentati in figura 5.3. Tenendo in considerazione entrambe le sessioni (figura 5.3c), si evidenzia una preferenza per SANKEY, soprattutto per quanto concerne l'efficacia di rappresentazione dei dati (Q1) e l'intuitività dell'interfaccia (Q2), mentre non si è evidenziata una preferenza tra SANKEY e CLAIRE per quanto riguarda la completezza del sistema per la risoluzione dei task (Q3).

Nella prima sessione CLAIRE è stato giudicato più completo di SANKEY ed è stato complessivamente preferito a quest'ultimo, nonostante SANKEY sia stato giudicato più intuitivo. Al contrario nella seconda sessione lo strumento SANKEY è stato giudicato preferibile per tutte le domande poste.

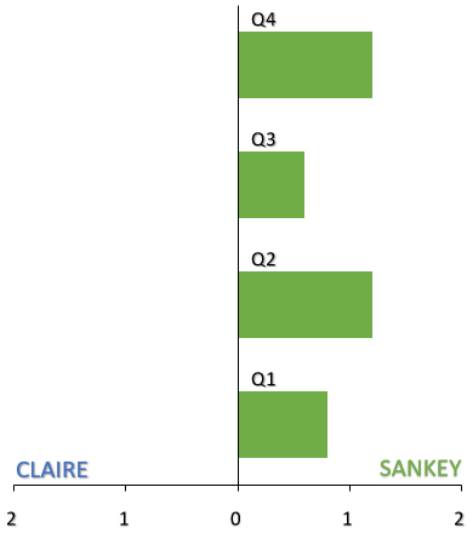
Nonostante SANKEY sia stato preferito da sei delle nove persone partecipanti al test, CLAIRE ha garantito il maggior numero di risposte corrette durante la risoluzione dei task (74% di risposte corrette con SANKEY, 89% di risposte corrette con CLAIRE). In generale sono state date risposte errate per entrambi i sistemi. I motivi di errore evidenziati sono principalmente due:

- interpretazione errata della richiesta del task;
- errore di interpretazione dei dati visualizzati.

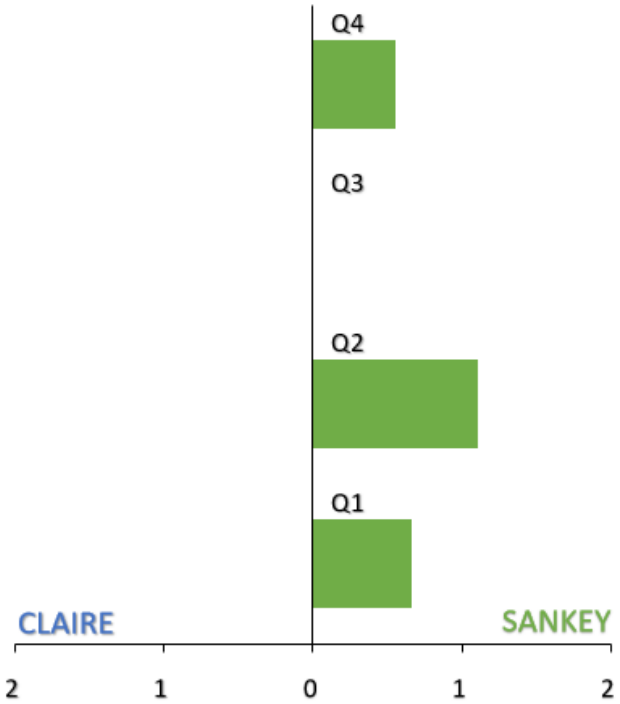
Per SANKEY la maggior parte degli errori si sono verificati per una interpretazione errata del posizionamento dei link finali. Molti utenti, infatti, hanno interpretato un ordinamento dei link incidenti ad uno stesso intervallo



(a) *Risultati ottenuti per il questionario di preferenza nella prima sessione di test (quattro utenti), dove è stato utilizzato prima il sistema SANKEY e successivamente CLAIRE.*



(b) *Risultati ottenuti per il questionario di preferenza nella seconda sessione di test (cinque utenti), dove è stato utilizzato prima il sistema CLAIRE e successivamente SANKEY.*



(c) *Risultati complessivi per il questionario di preferenza, tenendo in considerazione le risposte date da tutti e nove gli utenti partecipanti al test di usabilità.*

Figura 5.3. *Confronto dei due sistemi basato sul questionario di preferenza.*

di valori come indice di valutazione, aspettandosi che la posizione in cui un link è incidente desse indicazione circa il vero valore di performance ottenuto da un sistema di IR. Sebbene una categoria finale rappresenti un intervallo di valori, tale ordinamento non è presente nel sistema. Se due sistemi di IR $S1$ ed $S2$ ottengono prestazioni appartenenti ad uno stesso intervallo (ad esempio l'intervallo $[0.20, 0.24)$), con $S1$ che ottiene performance superiori di $S2$ (ad esempio $S1 = 0,22$ ed $S2 = 0,21$), non è detto che il link finale associato al sistema $S1$ sia incidente al nodo in una posizione inferiore (l'ordinamento degli intervalli da 0 a 1 è dall'alto verso il basso) rispetto alla posizione di incidenza del link finale associato al sistema $S2$.

Nonostante i questionari abbiano dato indicazioni circa l'utilità di entrambi gli strumenti visuali, sono stati raccolti dei commenti degli utenti che hanno partcipato al test di usabilità. Questa fase è stata fatta in particolare per lo strumento visuale SANKEY a seguito degli ulteriori cinque task assegnati. In generale i dati del questionario vengono parzialmente confermati dai commenti degli utenti. Un utente scrive: *"In una fase iniziale SANKEY è risultato essere particolarmente semplice da utilizzare, tuttavia CLAIRE, dopo un po' di allenamento, è sembrato più completo"*. Tale parere è stato espresso anche da un altro utente: *"SANKEY è più facile da utilizzare, CLAIRE invece è più potente perché permette di evidenziare immediatamente i sistemi più promettenti e quelli che ottengono le migliori performance, ma il confronto tra questi non sempre risulta immediato"*. Ciò viene particolarmente evidenziato nei risultati ottenuti dai questionari nella prima sessione di test, quando SANKEY è stato impiegato prima di CLAIRE. Un utente del secondo gruppo ha invece affermato che *"il sistema SANKEY è risultato più intuitivo solo grazie alla fase di spiegazione iniziale di entrambi i sistemi, ma se avessi dovuto provare ad utilizzarli entrambi senza alcuna spiegazione, probabilmente avrei riscontrato maggiori difficoltà ad utilizzare SANKEY rispetto a CLAIRE"*. Questo sta ad indicare come l'utilizzo di un diagramma di Sankey per questi tipi di rappresentazione, in cui si intendono evidenziare relazioni e combinazioni tra componenti differenti, è originale, di conseguenza può risultare poco comprensibile senza una fase di iniziale

di spiegazione. Molti utenti hanno accolto con entusiasmo questo tipo di rappresentazione indicandola come *"efficace, soprattutto perché, rispetto alla rappresentazione matriciale di CLAIRE, evidenzia meglio il concetto di sistema di IR come combinazione di componenti differenti"*.

Sono stati dati diversi suggerimenti di miglioramento per il sistema SANKEY. Un utente ha apprezzato molto l'idea di dimensionamento dinamico dei nodi basato sui valori medi di performance, ma ha evidenziato come in alcuni casi la differenza di dimensione risulti poco evidente e ha quindi proposto di lavorare su questo aspetto. Lo stesso utente ha anche auspicato un ordinamento iniziale di ciascuna famiglia di nodi in base al colore: *"questo permetterebbe di evidenziare meglio le differenze di colore tra due componenti appartenenti alla stessa famiglia e questo può risultare utile, soprattutto se i colori assumono un significato specifico come nel caso di stoplist e stemmer"*. Un altro utente ha invece proposto un ordinamento dinamico dei componenti in base alla dimensione, in modo da evideziare immediatamente quale componente ottiene in media performance minori o maggiori delle altre. Lo stesso utente ha anche evidenziato difficoltà di distinzione dei link finali e una conseguente difficoltà nel visualizzare il tooltip associato. Sebbene la funzione di scaling *min-max* in molte situazioni aiuti ad aumentare la distinzione dei link, l'utente ha proposto di permettere la scelta di visualizzazione di una sola delle categorie finali (ovvero di un solo intervallo di valori). Il nodo finale al quale l'utente è interessato potrebbe in questo modo sfruttare interamente lo spazio in altezza (e quindi "allargarsi") per permettere una maggior distinzione dei link incidenti ad esso.

La fase di valutazione ha permesso di riscontrare come il sistema SANKEY risulti essere particolarmente intuitivo nell'utilizzo (soprattutto dopo una veloce fase di spiegazione del sistema) ed efficace nel rappresentare i dati. Queste caratteristiche risultano evidenti anche se il sistema viene messo in comparazione con CLAIRE, il quale richiede una fase di addestramento più lunga, ma per alcuni utenti risulta particolarmente potente e completo.

6 Related works

Il sistema di Information Visualization (IV) proposto utilizza un diagramma di Sankey per la rappresentazione dei dati di valutazione dei sistemi di reperimento dell'informazione. Tali diagrammi sono generalmente impiegati per rappresentare un trasferimento di flusso da un insieme di nodi ad un altro, dove ciascun trasferimento viene rappresentato tramite un link la cui dimensione è proporzionale alla quantità di flusso. I diagrammi di Sankey prendono il nome dal capitano irlandese Matthew Henry Phineas Riall Sankey che per primo ha utilizzato questo tipo di schema per raffigurare l'efficienza energetica di un motore a vapore [Sankey, 1896]. Sankey prese spunto dal lavoro di Charles Joseph Minard che, nel 1861, disegnò un diagramma di flusso in combinazione con dati geografici e temporali, per rappresentare le pesanti perdite di uomini dell'esercito francese durante la campagna di Russia di Napoleone, del 1812 [Tufte, 1986]. Il diagramma rappresenta, oltre al percorso seguito dall'esercito di Napoleone, la variazione di dimensione dell'esercito a partire dal confine tra Russia e Polonia, fino all'arrivo a Mosca e ritorno. Tale rappresentazione permette di evidenziare come dei 422.000 uomini dai quali l'esercito era inizialmente composto, ne hanno fatto ritorno soltanto 10.000.

I diagrammi di Sankey vengono attualmente utilizzati in diversi campi. Oltre che per rappresentare trasferimenti di materiali e costi, vengono applicati, ad esempio, per la visualizzazione dei consumi energetici dei vari paesi del

mondo[1] o per la visualizzazione del traffico di un sito web. Google sfrutta i diagrammi di Sankey nel suo servizio Google Analytics per permettere agli utenti di analizzare in modo approfondito la variazione di flusso dei visitatori e la loro provenienza. La modalità di lettura dei dati *Flow Visualization* sfrutta diagrammi di Sankey per analizzare in maniera approfondita come gli utenti utilizzano il sito web, come si muovono e soprattutto da dove provengono. I nodi di un diagramma sono generati automaticamente attraverso un algoritmo che raggruppa i flussi più comuni di visitatori attraverso il sito. *Goal Flow* evidenzia il numero di utenti che hanno portato a compimento uno specifico percorso del sito. Questo permette di evidenziare gruppi di pagine interessanti del sito web oppure situazioni in cui il traffico agisce in modo inaspettato, permettendo di rilevare elementi che non vengono correttamente rappresentati in fase di navigazione [Mui, 2011]. I diagrammi di Sankey non sono mai stati utilizzati nel campo del reperimento dell'informazione.

In generale gli strumenti di visualizzazione proposti nel campo del reperimento dell'informazione hanno lo scopo di facilitare la presentazione e l'esplorazione dei documenti gestiti da un motore di ricerca. In [Fowler et al., 1991] viene presentato un sistema che utilizza tecniche di visualizzazione per la rappresentazione di una query, del contenuto dei documenti e delle relazioni tra termini. Il sistema permette la modifica di una query attraverso una manipolazione diretta della sua forma visuale e la visualizzazione dei documenti recuperati attraverso una mappa concettuale. In [Morse et al., 2002] vengono presentati diversi metodi di visualizzazione e rappresentazione dei dati, proposti per facilitare l'utente nel recupero dei documenti di interesse. Il sistema VIBE, ad esempio, rappresenta i termini di una query (detti *punti d'interesse*) come vertici di una figura, i documenti recuperati vengono raffigurati con delle icone sparse all'interno della figura, il loro posizionamento e la distanza dai punti di interesse dà una indicazione circa il contenuto del documento [Olsen et al., 1993]. In [Lipani et al., 2017] viene presentata una tecnica di visualizzazione per il metodo di pooling. Lo strumento di IV *Visual*

[1] http://www.iea.org/Sankey/

Pool permette all'utente di interagire facilmente con il metodo di pooling, integrando alcuni suggerimenti per analizzare le collezioni sperimentali ed aiutare l'utente a costruire migliori tecniche di pooling.

Questi approcci,tuttavia, non propongono tecniche di visualizzazione per facilitare la valutazione sperimentale. A tal proposito vengono presentati in dettaglio VIRTUE, uno strumento visuale che permette di effettuare analisi delle performance e failure analysis [Angelini et al., 2014], VATE^2, che facilita il processo di valutazione sperimentale introducendo la *what-if analysis* [Angelini et al., 2016], e CLAIRE, lo strumento utilizzato come base di comparazione per il sistema SANKEY, che consente di esplorare le prestazioni ottenute da una grande quantità di sistemi di IR [Angelini et al., 2017].

6.1 VIRTUE

In [Angelini et al., 2014] viene presentato *VIRTUE* (Visual Information Retrieval Tool for Upfront Evaluation), uno strumento visuale sviluppato con lo scopo di facilitare e rendere più efficiente il processo di valutazione dei sistemi di reperimento dell'informazione. VIRTUE supporta l'*analisi delle performance* e la *failure analysis* permettendo all'utente di analizzare ed esplorare le performance ottenute da alcuni sistemi e scoprire regioni critiche nel ranking dei documenti. L'analisi delle performance viene ottenuta tramite la visualizzazione interattiva delle curve appartenenti alla famiglia delle metriche *Cumulative Gain* (figura 6.1), ovvero CG, DCG, nCG e nDCG. VIRTUE permette l'analisi a due livelli:

1. l'analisi a livello di singolo topic (*topic level*) dove oltre alla curva della metrica selezionata vengono visualizzate altre due curve comparative: l'*ideal curve*, la curva ottenuta da una run ideale per lo specifico topic, e l'*optimal curve*, la curva ottenuta nel caso ottimo, ovvero dal riordinamento migliore possibile dei risultati rilevanti della run sperimentale;

2. l'analisi su un insieme di topic della collezione considerata (*experiment level*)
   per permettere all'utente di comprendere le prestazioni complessive del sistema. Per ciascun tipo di run (sperimentale, ideale e ottima) vengono mostrate cinque curve, che rappresentano il limite superiore, il quartile superiore, la mediana, il quartile inferiore e il limite inferiore.

L'obiettivo dell'analisi delle performance è quello di fornire semplici mezzi visuali per poter osservare se un sistema ha il potenziale per raggiungere la performance migliori oppure se è preferibile utilizzare una nuova strategia di ranking.

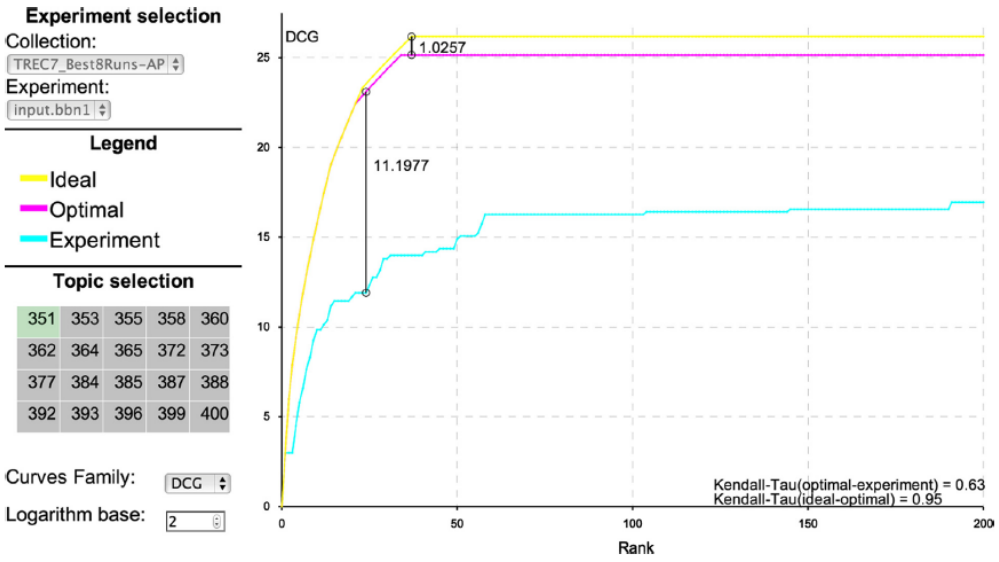


Figura 6.1. *Esempio di analisi a livello di singolo topic con VIRTUE [Angelini et al., 2014].*

VIRTUE consente di effettuare una failure analysis grazie all'utilizzo di due indicatori, la *Relative Position* (RP) e il *Delta Gain* (ΔG), che permettono visivamente di determinare i punti deboli e i punti forti del ranking. Anche in questo caso si identificano due situazioni per l'analisi:

1. la *failing documents identification* consente di scoprire quali documenti sono stati posizionati erroneamente nel ranking rispetto al caso ideale. Questo risulta possibile grazie alle curve della metrica selezionata,

associate con due grafici a barre colorate che rappresentano la *RP bar* e la ΔG *bar*. La RP bar, in base al colore visualizzato, determina se un documento è in posizione corretta rispetto al raking ideale oppure se è stato posizionato in un rank superiore o inferiore. La ΔG bar rappresenta per ogni documento mal posizionato la relazione tra il suo rank e il guadagno o la perdita procurata alla metrica considerata;

2. la *failing topics identification* consente di evidenziare il contributo generato da documenti posizionati erroneamente nel ranking rispetto al caso ideale per un certo insieme di topic.

La failure analysis è una attività fondamentale, ma che normalmente non viene eseguita in quanto molto impegnativa. VIRTUE assiste l'utente nella pratica della valutazione facilitando l'interpretazione delle curve CG-DCG e l'interazione con esse ed evidenziando aree critiche nel ranking dei documenti fornendo all'utente gli strumenti per rilevare le cause del fallimento. VIRTUE tuttavia non permette di effettuare un confronto diretto tra vari sistemi di IR ed in particolare non permette di ispezionare i singoli componenti di un sistema di reperimento, ma tratta l'intero sistema come una *black box*, interessandosi all'output delle metriche considerate e al ranking dei documenti.

6.2 VATE^2

In [Angelini et al., 2016] viene presentato VATE^2 (Visual Analytics Tool for Experimental Evaluation), una evoluzione di VIRTUE, in cui viene introdotto e sviluppato il concetto di *what-if analysis* (figura 6.2). La what-if analysis viene effettuata con l'obiettivo valutare quali sono gli effetti generati da una modifica di un sistema di IR prima che questa venga applicata. Questo tipo di valutazione risulta possibile grazie a tecniche di Visual Analytics (VA) presenti nel sistema VATE^2 che permettono:

- di esplorare la lista ordinata di risultati prodotti da un sistema di IR e le sue performance ottenute;

- di ipotizzare le possibili cause di documenti mal posizionati nella lista e quindi di individuare possibili correzioni;

- di stimare il possibile impatto generato dalle correzioni individuate.

Il sistema, esattamente come VIRTUE, supporta l'*analisi delle performance*, effettuata attraverso la visualizzazione delle curve appartenenti alla famiglia delle metriche Cumulative Gain, e la *failure analysis*, grazie all'utilizzo di un grafico a barre colorate che rappresenta la *RP bar* (Relative Position).

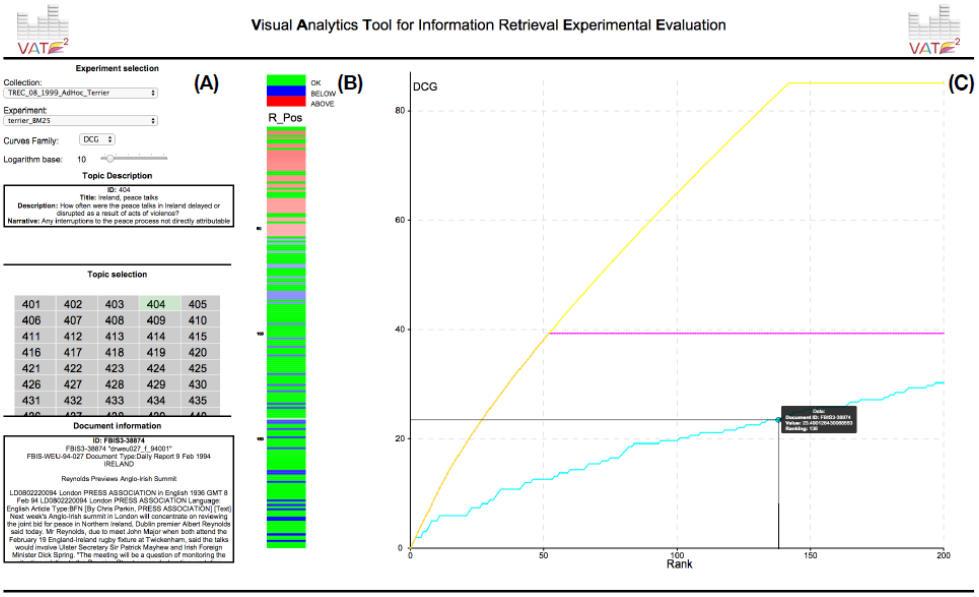


Figura 6.2. *Visualizzazione complessiva del sistema VATE [Angelini et al., 2016].*

In aggiunta VATE^2 permette, attraverso una funzione di drag & drop, di spostare un documento in una nuova posizione della lista ordinata. L'utente può individuare gli effetti generati sugli altri documenti della lista grazie alla possibilità di confrontare la nuova lista con l'originale. L'effetto dello spostamento viene visualizzato anche tramite le curve che rappresentano le performance ottenute dal sistema. La nuova curva (presentata con una linea continua) è confrontabile con la curva ottenuta prima dello spostamento effetuato (linea tratteggiata).

Grazie a VATE^2 è possibile stimare gli effetti generati dalla modifica di un sistema di IR al fine di migliorarlo, ma non permette di confrontare le performance ottenute da vari sistemi.

6.3 CLAIRE

CLAIRE (A Combinatorial Visual Analytics System for Information Retrieval Evaluation) è uno strumento visuale progettato per l'analisi e il confronto di un gran numero di sistemi di Information Retrieval [Angelini et al., 2017]. È lo strumento le cui funzionalità più si avvicinano a quelle del sistema SANKEY proposto in questa tesi, per questo motivo CLAIRE è stato utilizzato nel capitolo 5 come base di confronto per la validazione.

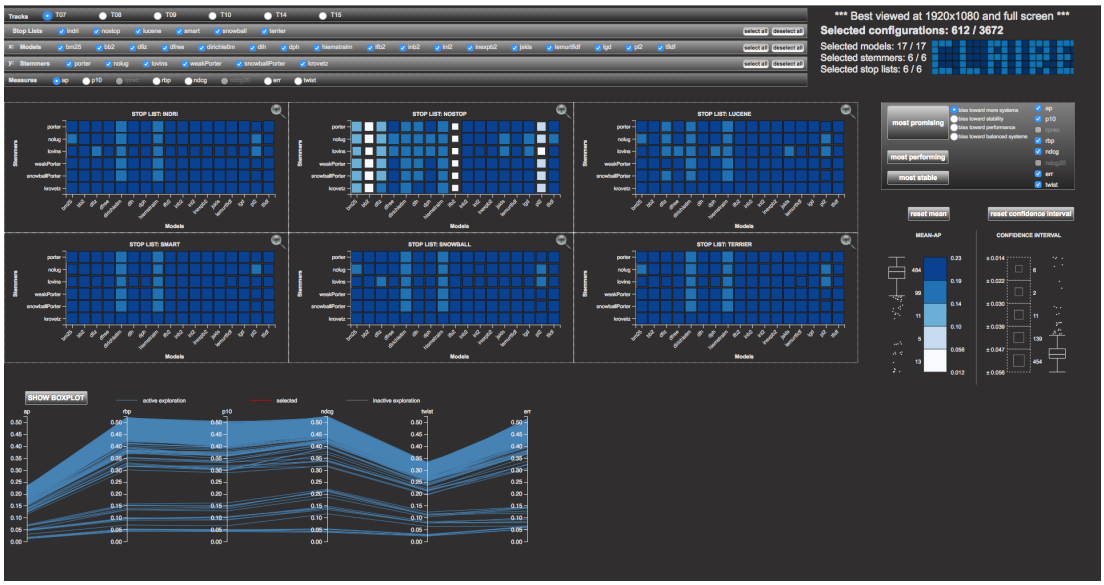


Figura 6.3. *Screenshot di Claire.*

CLAIRE è strutturalmente diviso in tre sezioni (figura 6.3):

1. l'area di *selezione dei parametri* per permettere all'utente di interagire con l'esplorazione dei sistemi, consentendogli di scegliere quale collezione di documenti considerare, ma anche quali stoplist, stemmer e modelli visualizzare e quale metrica di valutazione utilizzare;

2. l'area di *analisi dei sistemi*, che visualizza i dati basandosi sui parametri scelti dall'utente grazie ad una particolare tecnica visuale. Consente all'utente di effettuare un'analisi delle performance per la metrica scelta, su tutti i sistemi di IR considerati;

3. l'area di *valutazione complessiva* che fornisce all'utente un utile strumento basato su un grafico a coordinate parallele per osservare come i sistemi si comportano per l'intero insieme delle metriche disponibili.

CLAIRE visualizza i dati utilizzando una sequenza di *tile*, ciascuna contenente una matrice bidimensionale, in cui ogni quadrato rappresenta la performance di un determinato sistema (figura 6.4). In particolare il sistema viene identificato dai corrispondenti valori degli assi x e y e dal componente assegnato alla tile. La configurazione (ovvero i tipi di componenti assegnati alle tile e agli assi) può essere cambiata dall'utente utilizzando una funzione di drag and drop sulle famiglie di componenti nell'area di selezione dei parametri.

Figura 6.4. *Dettaglio di Claire. In questa sezione sono rappresentati tutti i dati di valutazione dei sistemi che utilizzano la stoplist indri. Un sistema selezionato viene evidenziato da un contorno rosso. Nell'esempio è selezionato il sistema che utilizza stemmer lovins e modello dlh.*

La scala di colori utilizzata per i quadrati varia dal bianco al blu scuro, dove il colore bianco identifica un valore basso della metrica considerata, il

blu scuro invece un valore alto. Le dimensioni dei quadrati rappresentano le dimensioni degli intervalli di confidenza, dove dimensioni minori indicano valori più piccoli. Entrambe queste informazioni sono espresse utilizzando una scala discreta a cinque valori, analizzabili grazie alla presenza di una legenda. Quindi CLAIRE, a differenza di SANKEY, utilizza un layout simile ad una matrice, il quale, oltre a garantire una visione compatta di tutte le differenti configurazioni, permette di evidenziare efficacemente eventuali particolarità nei dati. Oltre a ciò CLAIRE, tramite un grafico a coordinate parallele, garantisce per i sistemi visualizzati una visione d'insieme su tutte le altre metriche di valutazione considerate. Inoltre permette di evidenziare in maniera diretta i sistemi *più promettenti* (quelli con alte performance e bassa variabilità), i *più performanti* (cioè quelli con alte performance) o i *più stabili* (i sistemi con intervalli di confidenza più piccoli). Le analisi condotte e presentate nel capitolo 5 hanno permesso di evidenziare come l'utilizzo di CLAIRE non sempre risulti intuitivo e in generale la rappresentazione tramite matrici ad un primo approccio è risultata per molti più difficoltosa da comprendere rispetto alla rappresentazione proposta dal sistema SANKEY.

7 Conclusioni

Il reperimento dell'informazione si concentra sull'analisi, la rappresentazione, la memorizzazione, l'accesso e il reperimento dell'informazione contenuta in una collezione di documenti. Un sistema di IR ha l'obiettivo di recuperare documenti che siano in grado di soddisfare una esigenza informativa espressa da un utente tramite una interrogazione che viene generalmente posta al sistema attraverso l'utilizzo di alcune parole chiave. Per comprendere quanto bene un sistema di IR si comporti è necessario valutarlo e questo viene fatto attraverso l'utilizzo di collezioni sperimentali. La valutazione dei sistemi di reperimento dell'informazione ha lo scopo di valutare la loro efficacia nel recupero di documenti rilevanti a fronte dell'esigenza informativa dell'utente. Tuttavia la valutazione sperimentale giudica un sistema nella sua interezza, non fornendo alcuna informazione circa le performance di un singolo componente del sistema di IR o le interazioni tra componenti (stop list, stemmer, modelli). A tal fine, è opportuno sperimentare tutte le possibili combinazioni dei componenti disponibili per poi esplorare tutti i dati di valutazione generati dai vari sistemi di IR. Per individuare eventuali particolarità dovute alle interazioni dei componenti è necessario l'utilizzo di strumenti d'analisi manuale molto complessi.

In questa tesi è stato presentato uno strumento di Information Visualization, SANKEY, che facilita l'utente nell'esplorazione di un grande numero di dati di valutazione di sistemi di reperimento dell'informazione.

In particolare SANKEY viene utilizzato per la rappresentazione di più di un milione di punti di dati di valutazione (la Grid of Points), generati considerando sistemi nati dalla combinazione di sei stoplist, sei stemmer e diciassette modelli di IR, che sono stati valutati su sei diverse collezioni sperimentali (ciascuna comprendente cinquanta topic) utilizzando sei metriche di valutazione. SANKEY, tramite l'utilizzo di una rappresentazione che prende spunto dai diagrammi di Sankey e utili strumenti di Information Visualization e Visual Analytics, facilita l'esplorazione dei dati, evitando all'utente un'analisi manuale che sarebbe troppo onerosa e che richiederebbe l'utilizzo di strumenti statistici complessi per la comprensione degli effetti dati da un singolo componente o una specifica interazione tra questi.

Il sistema SANKEY utilizza una semplice interfaccia suddivisa in due principali sezioni. La prima sezione di selezione dei parametri permette all'utente di interagire con la scelta dei dati da visualizzare. È possibile selezionare:

- la collezione sperimentale da analizzare;
- uno specifico topic della collezione (se si vuole effettuare un'analisi più specifica);
- la metrica di valutazione da considerare;
- i singoli componenti da visualizzare per ciascuna famiglia (stoplist, stemmer, modelli);

Tale sezione permette anche di invertire il posizionamento delle componenti che formano un sistema di IR per poter evidenziare al meglio alcune particolarità dei dati in fase di analisi ed esplorazione visuale. Inoltre è presente una legenda dei colori per aiutare l'utente a comprendere alcune informazioni associate alle componenti, quali: la tipologia del modello di IR (probabilistico, vettoriale, language model), il numero di stopword in una stoplist, il livello di aggressività di uno stemmer.

La seconda sezione presenta il diagramma di Sankey generato per rappresentare i sistemi di IR e la loro valutazione. Ciascun diverso componente viene visualizzato attraverso un nodo del diagramma, mentre un link che connette due componenti rappresenta una combinazione o interazione tra componenti. Un sistema di IR viene rappresentato da un percorso che connette una stoplist, uno stemmer, un modello e un valore finale rappresentante la categoria che identifica la performance ottenuta dal sistema. L'utente può interagire con il sistema, ad esempio:

- visualizzando, attraverso dei tooltip, alcune informazioni aggiuntive associate a nodi e link;
- selezionando alcuni dei nodi per evidenziare i percorsi rappresentanti tutti i sistemi ai quali è interessato;

Il sistema supporta l'utente nell'esplorazione dei dati anche attraverso alcune tecniche di visual analytics:

- ciascun nodo rappresenta l'insieme di sistemi di IR che utilizzano lo specifico componente, il dimensionamento del nodo è proporzionale alla media aritmetica delle performance ottenute da questi sistemi;
- ciascun link che connette due nodi associati a due componenti rappresenta l'insieme di sistemi di IR che utilizzano quei componenti, il dimensionamento in larghezza del link è proporzionale alla media aritmetica delle performance ottenute da questi sistemi;
- si applica per ciascun insieme di sistemi di IR (associati ad un nodo o un link) il test statistico di Dunnett per evidenziare il top group, ovvero il gruppo formato dai sistemi che non si differenziano significativamente dal miglior sistema tra quelli considerati.

È stata condotta una fase di validazione dello strumento SANKEY attraverso un test di usabilità. Nove utenti hanno partecipato al test che ha

messo a confronto il sistema SANKEY con un altro strumento visuale per l'esplorazione e l'analisi dei dati di valutazione dei sistemi di IR, chiamato CLAIRE. Il test di usabilità, a seguito della risoluzione di tre task, ha presentato due questionari: uno di valutazione dei due singoli sistemi e uno di preferenza tra i due sistemi di Information Visualization. I risultati del primo questionario hanno permesso di evidenziare come SANKEY risulti essere particolarmente intuitivo nell'utilizzo ed efficace nella rappresentazione dei dati. Il secondo questionario ha evidenziato una preferenza generale per il sistema SANKEY, tuttavia entrambi i sistemi si sono rivelati particolarmente completi per la risoluzione dei task proposti, anche se il sistema CLAIRE ha permesso agli utenti di risolvere correttamente l'89% dei task, contro il 74% del sistema SANKEY. È stato riscontrato come gli errori effettuati dagli utenti siano stati causati da una cattiva interpretazione delle richieste del task o dei dati visualizzati. Agli utenti è stato richiesto di effettuare cinque ulteriori task utilizzando il sistema SANKEY per permettere di esprimere una opinione più approfondita dello strumento. Questo ha permesso di confermare i risultati ricavati dai questionari, con il sistema SANKEY che è ritenuto dagli utenti più intuitivo e facile da utilizzare, mentre CLAIRE è stato considerato più completo e potente. Per quanto riguarda gli sviluppi futuri, l'obiettivo principale è quello di migliorare il sistema cercando di renderlo ancora più intuitivo, semplice da utilizzare e completo. In fase di test sono stati accolti diversi suggerimenti di miglioramento, ad esempio la possibilità di ordinare in maniera automatica i nodi all'interno di una famiglia in base a vari criteri di ordinamento (colori o dimensioni).

Inoltre il sistema SANKEY è stato applicato su dati generati per collezioni sperimentali in lingua inglese. Un'altro obiettivo è quello di allargare le possibilità di scelta per l'utente, permettendogli di analizzare sistemi di IR su altre collezioni sperimentali di lingue differenti. Questo richiederebbe di gestire svariati nuovi sistemi, creati da combinazioni di componenti diversi (stoplist e stemmer dipendono dalla lingua della collezione).

Bibliografia

[Amati and Van Rijsbergen, 2002] Amati, G. and Van Rijsbergen, C. J. (2002). Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. Inf. Syst.*, 20(4):357–389.

[Angelini et al., 2017] Angelini, M., Fazzini, V., Ferro, N., Santucci, G., and Silvello, G. (2017). CLAIRE: a combinatorial visual analytics system for information retrieval evaluation. *submitted to: IEEE Transactions on Visualization and Computer Graphics - under review*.

[Angelini et al., 2014] Angelini, M., Ferro, N., Santucci, G., and Silvello, G. (2014). VIRTUE: A visual tool for information retrieval performance evaluation and failure analysis. *Journal of Visual Languages and Computing*, 25(4):394–413.

[Angelini et al., 2016] Angelini, M., Ferro, N., Santucci, G., and Silvello, G. (2016). A visual analytics approach for what-if analysis of information retrieval systems. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, pages 1081–1084, New York, NY, USA. ACM.

[Baeza-Yates and Ribeiro-Neto, 1999] Baeza-Yates, R. A. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

[Croft et al., 2009] Croft, B., Metzler, D., and Strohman, T. (2009). *Search Engines: Information Retrieval in Practice*. Addison-Wesley Publishing Company, USA, 1st edition.

[Dunnett, 1955] Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association*, 50(272):1096–1121.

[Ferro and Silvello, 2016] Ferro, N. and Silvello, G. (2016). A general linear mixed models approach to study system component effects. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, pages 25–34, New York, NY, USA. ACM.

[Ferro and Silvello, 2017] Ferro, N. and Silvello, G. (2017). Towards an anatomy of ir system component performances. *Journal of the Association for Information Science and Technology (JASIST) - accepted for pubblications*.

[Ferro et al., 2016] Ferro, N., Silvello, G., Keskustalo, H., Pirkola, A., and Järvelin, K. (2016). The twist measure for IR evaluation: Taking user's effort into account. *Journal of the Association for Information Science & Technology*, 67(3):620–648.

[Fowler et al., 1991] Fowler, R. H., Fowler, W. A. L., and Wilson, B. A. (1991). Integrating query thesaurus, and documents through a common visual representation. In *Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '91, pages 142–151, New York, NY, USA. ACM.

[Harman, 2011] Harman, D. (2011). *Information Retrieval Evaluation*. Morgan & Claypool Publishers, 1st edition.

[Hiemstra, 1998] Hiemstra, D. (1998). *A Linguistically Motivated Probabilistic Model of Information Retrieval*, pages 569–584. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Krovetz, 1993] Krovetz, R. (1993). Viewing morphology as an inference process. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '93, pages 191–202, New York, NY, USA. ACM.

[Lipani et al., 2017] Lipani, A., Lupu, M., and Hanbury, A. (2017). Visual pool: A tool to visualize and interact with the pooling method. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, pages 1321–1324, New York, NY, USA. ACM.

[Lovins, 1968] Lovins, J. B. (1968). Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, 11:22–31.

[Manning et al., 2008] Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

[Morse et al., 2002] Morse, E., Lewis, M., and Olsen, K. A. (2002). Testing visual information retrieval methodologies case study: Comparative analysis of textual, icon, graphical, and "spring" displays. *Journal of the American Society for Information Science and Technology*, 53(1):28–40.

[Mui, 2011] Mui, P. (2011). Introducing flow visualization: visualizing visitor flow. `https://analytics.googleblog.com/2011/10/introducing-flow-visualization.html`. [Last accessed on 18 Settembre 2017].

[Nielsen and Landauer, 1993] Nielsen, J. and Landauer, T. K. (1993). A mathematical model of the finding of usability problems. In *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems*, CHI '93, pages 206–213, New York, NY, USA. ACM.

[Olsen et al., 1993] Olsen, K. A., Korfhage, R. R., Sochats, K. M., Spring, M. B., and Williams, J. G. (1993). Visualization of a document collection: The VIBE system. *Information Processing and Management*, 29(1):69–81.

[Porter, 1980] Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 14(3):130–137.

[Robertson and Jones, 1976] Robertson, S. E. and Jones, S. K. (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27(3):129–146.

[Salton et al., 1975] Salton, G., Wong, A., and Yang, C. S. (1975). A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620.

[Sankey, 1896] Sankey, H. R. (1896). The thermal efficiency of steam-engines. (including appendixes). *Minutes of the Proceedings of the Institution of Civil Engineers*, 125(1896):182–212.

[Tufte, 1986] Tufte, E. R. (1986). *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT, USA.

[van Rijsbergen and Jones, 1973] van Rijsbergen, C. J. and Jones, S. K. (1973). A test for the separation of relevant and non-relevant documents in experimental retrieval collections. *Journal of Documentation*, 29:251–257.

[Voorhees, 2001] Voorhees, E. M. (2001). Overview of TREC 2001. In *Proceedings of The Tenth Text REtrieval Conference, TREC 2001, Gaithersburg, Maryland, USA, November 13-16, 2001*.

[Voorhees, 2005] Voorhees, E. M. (2005). Overview of TREC 2005. In *Proceedings of the Fourteenth Text REtrieval Conference, TREC 2005, Gaithersburg, Maryland, USA, November 15-18, 2005*.

[Voorhees, 2006] Voorhees, E. M. (2006). Overview of the TREC 2006. In *Proceedings of the Fifteenth Text REtrieval Conference, TREC 2006, Gaithersburg, Maryland, USA, November 14-17, 2006*.

[Voorhees and Harman, 1999] Voorhees, E. M. and Harman, D. (1999). Overview of the eighth text retrieval conference (TREC-8). In *Proceedings of The Eighth Text REtrieval Conference, TREC 1999, Gaithersburg, Maryland, USA, November 17-19, 1999*.

[Voorhees and Harman, 2000] Voorhees, E. M. and Harman, D. (2000). Overview of the ninth text retrieval conference (TREC-9). In *Proceedings of The Ninth Text REtrieval Conference, TREC 2000, Gaithersburg, Maryland, USA, November 13-16, 2000*.

[Voorhees and Harman, 1998] Voorhees, E. M. and Harman, D. K., editors (1998). *Proceedings of The Seventh Text REtrieval Conference, TREC 1998, Gaithersburg, Maryland, USA, November 9-11, 1998*, volume Special Publication 500-242. National Institute of Standards and Technology (NIST).