Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN: 2548-964X
Vol. 4, No. 4, April 2020, hlm. 1290-1299 http://j-ptiik.ub.ac.id

Temu Kembali Informasi Lintas Bahasa Dokumen Berita Bahasa Indonesia-Inggris menggunakan Metode BM25F

Lusiyana Adetia Isadi[1], Indriati[2], Putra Pandu Adikara[3]

Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Brawijaya
Email: [1]lusiyana@student.ub.ac.id, [2]indriati.tif@ub.ac.id, [3]adikara.putra@ub.ac.id

Abstrak

Berita ialah sumber informasi yang ditampilkan pada khalayak umum mengenai suatu kejadian dan tersaji dalam berbagai macam bahasa. Umumnya pada suatu *website* hanya memungkinkan melakukan pencarian dalam satu bahasa saja. Hal ini menyebabkan masalah bagi para pengguna yang ingin lebih cepat menemukan informasi yang lebih luas dalam beberapa bahasa sekaligus. Masalah tersebut dapat diatasi dengan mengembangkan sistem temu kembali informasi lintas bahasa. Sistem tersebut dapat meningkatkan efisiensi waktu karena dapat mengembalikan dokumen dalam dua bahasa hanya dengan memasukkan *query* dalam satu bahasa saja. Salah satu metode yang dapat digunakan untuk mengembangkan sistem tersebut ialah metode BM25F yang dapat mengembalikan dokumen yang relevan serta menangani dokumen terstruktur. Struktur data berita yang digunakan pada pelatihan dan pengujian ialah bagian judul dan isi dari berita. Data yang digunakan ialah 300 dokumen berita bahasa Indonesia dan 300 dokumen berita bahasa Inggris kemudian dilakukan pengujian nilai *boost*, pengujian *query* bahasa Indonesia, dan pengujian *query* bahasa Inggris. Pada pengujian nilai *boost*, nilai *precision@k* tertinggi didapatkan pada saat *boost* judul bernilai 5 dan *boost* isi bernilai 1. Nilai tersebut akan digunakan untuk pengujian *query*. Pengujian *query* dilakukan dengan menggunakan *precision@k* dan mendapatkan nilai tertinggi pada *k*=5 yaitu sebesar 0,98 pada pengujian *query* bahasa Indonesia yang mengembalikan dokumen bahasa Indonesia dan Inggris.

Kata kunci: *berita, temu kembali informasi lintas bahasa, text mining, mesin pencarian, BM25F*

Abstract

News is a source of information that displayed to the general public about an event and presented in various languages. Usually, a website only allows user to search only in one language. This causes problems for users who want to find broader information more quickly in several languages at once. These problems can be overcome by developing a cross language information retrieval system. The system can improve the time efficiency because it can return documents in two languages by simply entering a query in one language only. One of the method that can be used to develop the system is BM25F method that can return relevant documents and handle structured documents. The news data structure used in training and testing is the title and the content part of the news. The data used in this study are 300 Indonesian news documents and 300 English news documents that will be used to test the boost value, the Indonesian queries, and the English queries. For the boost value testing, the highest precision@k value obtained when the title boost is 5 and the content boost is 1. This value will be used for query testing. Query testing is performed using precision@k and got the highest value of 0.98 when k=5 in the Indonesian queries test which returned Indonesian and English documents.

Keywords: *news, cross language information retrieval, text mining, search engine, BM25F*

1. PENDAHULUAN

Berita merupakan sumber informasi yang ditampilkan kepada khalayak umum mengenai suatu hal yang terjadi. Semakin berkembangnya teknologi, internet dijadikan sebagai salah satu media penyebaran berita antar negara yang sering digunakan. Berita yang tersebar di internet tersedia dalam berbagai macam bahasa, namun pada umumnya suatu *website* hanya memungkinkan satu bahasa untuk melakukan pencarian. Persentase jumlah pengguna yang

mengerti lebih dari satu bahasa cukup besar (Nie, 2010). Hal ini menyebabkan masalah bagi para pengguna tersebut yang ingin lebih cepat menemukan informasi yang lebih luas dalam beberapa bahasa sekaligus (Zhou, et al., 2012).

Informasi yang diperlukan pengguna mungkin tidak tersedia dalam bahasa asli yang digunakan olehnya. Informasi juga mungkin tersedia dalam campuran beberapa bahasa (Nie, 2010). Dengan mengembangkan sistem temu kembali informasi lintas bahasa atau yang biasa disebut *Cross Language Information Retrieval* (CLIR) bisa menjadi solusi untuk permasalahaan-permasalahan yang telah disebutkan sebelumnya.

CLIR berhubungan dengan pengambilan dokumen yang relevan dengan menggunakan *query* yang dituliskan dalam bahasa lain. Sistem CLIR menjadi sangat penting pada masa ini karena meningkatnya dokumen berita dengan berbagai macam bahasa (Elayeb & Bounhas, 2015). Penelitian mengenai CLIR umumnya berfokus pada metode terjemahan *query*. Hal ini disebabkan karena masalah komputasi yang lebih cepat (Zhou, et al., 2012).

Metode yang dapat digunakan untuk CLIR salah satunya ialah metode Best Match 25 (BM25). Metode BM25 dapat mengembalikan dokumen yang relevan terhadap *query* yang dicari dengan baik (Sari & Adriani, 2014). Metode BM25 tidak memperhatikan struktur dari dokumen pada proses pembobotannya. Sedangkan, dokumen berita merupakan dokumen terstruktur yang memiliki bagian judul dan isi berita. Maka untuk mengatasi masalah tersebut Robertson mengembangkan metode BM25 yang dinamakan metode BM25F yang dapat menangani dokumen terstruktur (Garcia, 2011).

Penelitian sebelumnya mengenai CLIR menggunakan dokumen bahasa Indonesia-Inggris dengan membandingkan metode BM25 dan BM25 Modifikasi. Data dalam penelitian tersebut ialah 600 data berbentuk berita dari situs Edition.cnn.com, Nytimes.com, serta Detik.com. Pengujian dilakukan dengan menggunakan *Precision@k* memberikan hasil tertinggi senilai 0,95 saat k=5 untuk setiap metodenya (Iriani, et al., 2019).

Salah satu penelitian yang menggunakan metode BM25F yaitu untuk pencarian semantik pada halaman web DBpedia. Hasil akurasi yang didapatkan dengan menggunakan *Mean Average Precision* (MAP) ialah 0,1411 untuk Lucene, 0,1243 untuk LuceneF, 0,1659 untuk BM25 dan 0,1743 untuk BM25F. Maka dapat diketahui bahwa BM25F memberikan hasil yang paling baik dibandingkan metode lainnya pada penelitian tersebut (Pérez-Agüera, et al., 2010).

Penelitian lainnya yang menggunakan metode BM25F ialah penelitian yang bertujuan untuk mengevaluasi *framework* untuk temu kembali dokumen *Extensible Markup Language* (XML). XML ialah bahasa *markup* untuk keperluan pertukaran data antar sistem. Pengujian yang dilakukan menggunakan *The Initiative for the Evaluation of XML retrieval* (INEX) 2009 *ad hoc task* untuk *retrieval* ad hoc dan INEX 2008 *Book Track data* untuk *retrieval* halaman buku. Dari hasil perbandingan metode BM25 dan BM25F, untuk *retrieval* ad hoc memberikan hasil akurasi sebesar 0,633 pada BM25F dan 0,594 pada BM25 dengan menggunakan *Interpolated Precision* atau *iP* [0,01]. Sedangkan, untuk *retrieval* halaman buku dengan pengujian MAP memberikan hasil akurasi sebesar 0,0278 (pelatihan) serta 0,0110 (pengujian) pada BM25 dan sebesar 0,0412 (pelatihan) serta 0,0149 (pengujian) pada BM25F judul (Itakura & Clarke, 2010).

Dari penelitian sebelumnya bisa disimpulkan jika metode BM25F bisa dipakai untuk melakukan pencarian dokumen serta untuk temu kembali informasi dengan hasil yang lebih baik dibandingkan metode lainnya pada dokumen yang terstruktur. Penelitian ini memiliki tujuan untuk menemukan dokumen yang relevan dari dokumen berita bahasa Indonesia-Inggris hanya dengan masukan *query* satu bahasa menggunakan metode BM25F.

2. DASAR TEORI

2.1. Berita

Berdasarkan Kamus Besar Bahasa Indonesia, berita ialah keterangan atau cerita terkait peristiwa atau kejadian yang hangat (KBBI, 2019). Suatu peristiwa dapat dianggap sebagai berita apabila peristiwa tersebut menarik, memiliki nilai penting, masih baru, aman saat disiarkan, dan mengandung nilai kebenaran (Wahyudi, 1991). Berita ialah kejadian hangat yang menarik dan disiarkan kepada publik.

2.2. Temu Kembali Informasi Lintas Bahasa

Temu Kembali Informasi Lintas Bahasa atau *Cross Language Information Retrieval* (CLIR) merupakan suatu cara untuk mencari

dokumen dalam bahasa yang berbeda dengan bahasa yang digunakan pada *query* (Wang & Oard, 2012). Teknik CLIR dibagi menjadi 4 kategori berdasarkan sumber terjemahannya, yaitu teknik CLIR berdasarkan kamus yang menggunakan kamus untuk mengembalikan informasi dengan bahasa yang berbeda dari *query*, teknik CLIR berdasarkan *parallel corpora* yang mengambil informasi dari teks yang sebelumnya telah diterjemahkan ke dalam dua atau beberapa bahasa, teknik CLIR berdasarkan *comparable corpora* yang menggunakan *corpora* dalam berbagai bahasa namun bukan terjemahan melainkan berkaitan dengan subjek yang sama, dan teknik CLIR berdasarkan mesin penerjemah (CLEF, 2003).

Terdapat tiga pendekatan umum untuk melakukan CLIR yaitu, menerjemahkan *query* agar sesuai dengan dokumen, menerjemahkan dokumen agar sesuai dengan *query*, dan menerjemahkan dokumen dan *query* menjadi bahasa ketiga (Zhou, et al., 2012). Pendekatan pada penelitian ini ialah pendekatan pertama, yaitu menerjemahkan *query*. Pendekatan ini dipilih karena masalah komputasi waktu yang lebih cepat dibandingkan pendekatan lainnya. Terjemahan *query* akan dilakukan oleh mesin penerjemah.

2.3. *Preprocessing*

Tahap penting yang dilakukan pada teks mining ialah *preprocessing*. *Preprocessing* dilakukan untuk mendapatkan term indeks dari dokumen ataupun *query* yang akan dilakukan untuk pembobotan. Proses pada tahap *preprocessing* antara lain *case folding*, *tokenizing*, *cleansing*, *stemming*, dan *stopword removal atau filtering* (Ganesan, 2019).

2.4. Metode BM25F

Metode BM25F merupakan algoritme ekstensi dari BM25. Algoritme BM25F berfungsi memberi peringkat untuk dokumen terstruktur (Pérez-Agüera, et al., 2010). Rumus dari BM25F (Pérez-Iglesias, et al., 2009) ditunjukkan pada Persamaan (1).

$$R(q,d) = \sum_{t\ in\ q} idf(t) \times \frac{weight(t,d)}{k_1 + weight(t,d)} \quad (1)$$

Keterangan:

- *R(q,d)* = *ranking* dari *query q* pada dokumen *d*.
- *idf(t)* = nilai *idf* yang ditunjukkan pada Persamaan (2).

$$idf(t) = log\frac{N - df(t) + 0.5}{df(t) + 0.5} \quad (2)$$

- N = jumlah seluruh dokumen dan *df(t)* merupakan jumlah dokumen yang memiliki term *t*.
- k_1 = konstanta yang mengatur pertumbuhan frekuensi term non-linier
- *weight(t,d)* = bobot term *t* pada dokumen *d* yang ditunjukkan pada Persamaan (3).

$$weight(t,d) = \sum_{c\ in\ d} \frac{occurs_{t,c}^{d} \times boost_c}{\left((1-b_c) + b_c \times \frac{l_c}{avl_c}\right)} \quad (3)$$

- *c* = *field* pada dokumen.
- *d* = dokumen.
- $occurs_{t,c}^{d}$ = kemunculan term *t* pada *field c* pada dokumen *d*.
- $boost_c$ = faktor *boost* yang diberikan pada *field c*.
- b_c = konstanta.
- l_c = panjang dari *field c*.
- avl_c = rata-rata dari panjang *field c*.

Menurut Robertson dan Walker (1999), nilai *default* dari k_1 dan b_c secara berturut-turut ialah 1,2 dan 0,75. Rumus IDF yang digunakan pada penelitian ini memungkinkan nilai negatif apabila nilai $df(t) > N/2$ (Fang, et al., 2004).

2.5. Evaluasi

Evaluasi dilakukan untuk mengukur kemampuan sistem yang dibuat. Pada penelitian ini metode evaluasi yang dipakai ialah evaluasi *Precision@K* yang dipakai untuk evaluasi temu kembali berperingkat. Evaluasi ini digunakan untuk penghitungan nilai persentase dokumen sebanyak *K* teratas dan mengabaikan dokumen yang hasil evaluasinya berada di bawah peringkat *K*.

Evaluasi ini melakukan pengukuran *precision* mulai dari level terendah dari seluruh hasil yang didapatkan, seperti 5, 10, sampai 30 dokumen. Keuntungan dari metode ini adalah tidak memerlukan ukuran dari dokumen yang relevan, tetapi kekurangannya ialah jumlah dokumen relevan untuk sebuah *query* memiliki pengaruh yang besar untuk *precision* pada *K* (Manning, et al., 2009). Rumus dari evaluasi *Precision@K* ditunjukkan pada Persamaan (4)

$$Precision@K = \frac{r}{K} \quad (4)$$

Keterangan:

- *r* = jumlah dokumen yang relevan terdapat pada *K* dokumen teratas.

- *K* = nilai batas peringkat.

3. METODOLOGI

3.1. Metode Pengumpulan Data

Pada penelitian ini, data yang digunakan ialah berita dari situs kompas.com dan thejakartapost.com. Data yang akan diproses berupa data berita bahasa Indonesia dari situs kompas.com dan data berita bahasa Inggris dari situs thejakartapost.com pada kategori *travel*. Dipilihnya kategori ini sebagai variable penelitian, selain untuk fokus penelitian yang lebih terarah, juga karena pada kategori ini terdapat sub-sub kategori yang dapat mendeskripsikan tema ini secara lebih luas namun dengan karakteristiknya masing-masing yang lebih spesifik, seperti "makan makan", "jalan jalan", "destinasi", dan sebagainya.

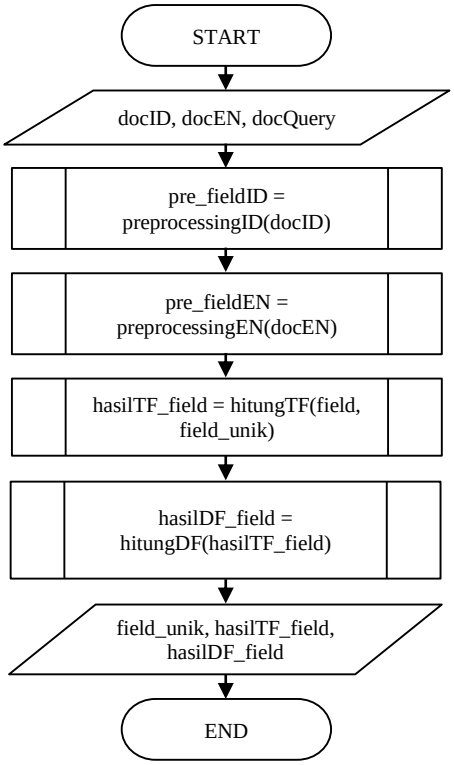
Jumlah data berita yang akan diambil pada masing-masing situs berita ialah sejumlah 300 data sehingga jumlah data yang digunakan pada penelitian ialah 600 data berita. Data berita yang digunakan diambil dari berita yang dipublikasikan pada Januari 2018 hingga Desember 2019. Pengambilan data dilakukan dengan cara disalin satu per satu secara manual ke dalam *file* Excel yang akan disimpan dalam format .csv.

3.2. Metode Analisa Data

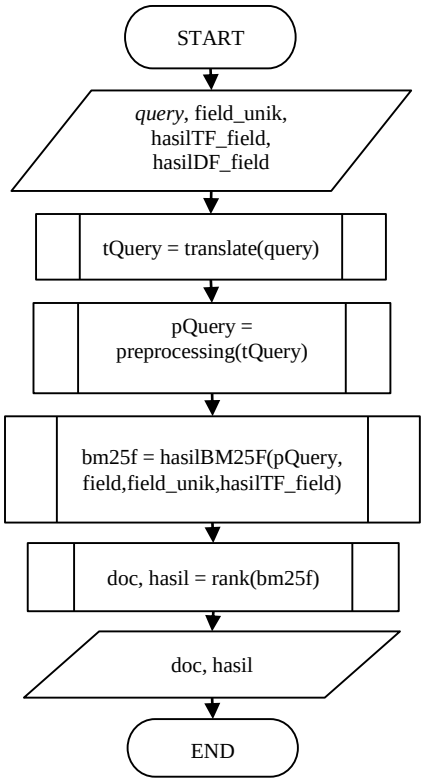
Pengujian yang dilakukan ialah pengujian *boost*, pengujian *query* bahasa Indonesia, dan pengujian *query* bahasa Inggris. Setiap pengujian akan dilakukan evaluasi dengan menggunakan *Precision@k*. Pada pengujian *boost* akan dilakukan pemilihan nilai yang terbaik, sedangkan untuk pengujian *query* akan dilakukan analisis penggunaan *query* yang mengembalikan hasil yang paling baik.

3.3. Diagram Alir Sistem

Metode yang digunakan untuk penelitian ini ialah metode BM25F. Terdapat dua tahap yang akan dilakukan, yaitu tahap pelatihan serta tahap pengujian. Pada tahap pelatihan, dilakukan *preprocessing* untuk semua dokumen berita bahasa Indonesia dan bahasa Inggris, kemudian setiap term yang ada di dokumen akan dihitung bobotnya dengan pembobotan TF. Pada tahap pengujian, masukan berupa *query* dan hasil pembobotan TF yang didapatkan dari tahap pelatihan. *Query* akan diterjemahkan lalu dilakukan tahap *preprocessing* untuk setiap bahasa. Dari hasil *preprocessing* tersebut akan dilakukan perhitungan BM25F dan mendapatkan hasil pemeringkatan berita. Diagram alir tahap pelatihan ditunjukkan pada Gambar 1. Diagram alir tahap pengujian ditunjukkan pada Gambar 2.



Gambar 1. Diagram Alir Tahap Pelatihan



Gambar 2. Diagram Alir Tahap Pengujian

4. PENGUJIAN DAN ANALISIS

Pengujian yang dilakukan pada penelitian ini antara lain pengujian nilai *boost*, pengujian *query* bahasa Indonesia, dan pengujian *query* bahasa Inggris. Sebelum melakukan pengujian nilai *boost*, akan dilakukan pengujian *precision@k* BM25F untuk mendapatkan relevan atau tidaknya suatu dokumen berita yang dikembalikan oleh sistem. Pada pengujian ini suatu query akan dimasukkan dan masing-masing query akan mengeluarkan maksimal 30 data berita. Semua data yang digunakan pada pengujian ini bisa dilihat di lampiran yang dapat diakses pada http://bit.ly/lampiran-lusi.

Query yang digunakan pada tahap pengujian berasal dari 1 *user*. *Query* yang digunakan berlandaskan dari penelitian yang dilakukan oleh Fesenmaier et al. (2010) yang memberikan pernyataan bahwa pada pencarian travel umumnya seseorang mencari informasi mengenai suatu destinasi, harga, tempat tinggal, transportasi, panduan, diskon dan promosi, apa yang harus dilakukan selama liburan, serta mengenai suatu acara tertentu.

4.1. Pengujian Nilai *Boost*

Pengujian pada subbab ini dilakukan demi mengetahui nilai *boost* yang paling baik untuk masing-masing *field*. Nilai *boost* merupakan nilai yang merepresentasikan seberapa pentingnya suatu *field* dokumen. Pada penelitian yang dilakukan oleh Robertson & Zaragoza (2009) nilai *boost* yang digunakan pada rumus BM25F ialah 1 sampai dengan 5. Nilai ini akan digunakan untuk mendapatkan nilai *boost* terbaik untuk masing-masing *field*. *Query* yang digunakan pada pengujian ini ditunjukkan pada Tabel 1.

Tabel 1. Daftar *Query* Pengujian *Boost*

No	Query
1	Rekomendasi destinasi wisata
2	Recommendation tourist destinations
3	Makanan khas Indonesia
4	Indonesia Traditional Foods

Hasil pengujian nilai *boost* ditunjukkan pada Tabel 2.

Tabel 2. Pengujian Nilai *Boost*

Q	Boost		*Precision@k* BM25F					
	Judul	Isi	*k=5*	*k=10*	*k=15*	*k=20*	*k=25*	*k=30*
1	1	1	1	0,9	0,8	0,85	0,84	0,833
	1	2	1	0,9	0,867	0,85	0,88	0,833
	1	3	0,8	0,9	0,867	0,85	0,84	0,833
	1	4	0,8	0,9	0,867	0,85	0,84	0,833
	1	5	0,8	0,9	0,867	0,85	0,84	0,833
	2	1	1	0,9	0,933	0,85	0,84	0,833
	2	2	1	0,9	0,8	0,85	0,84	0,833
	2	3	1	0,9	0,8	0,85	0,84	0,833
	2	4	1	0,9	0,867	0,85	0,88	0,833
	2	5	0,8	0,9	0,867	0,85	0,88	0,833
	3	1	1	0,9	0,933	0,85	0,88	0,867
	3	2	1	0,9	0,867	0,85	0,84	0,833
	3	3	1	0,9	0,8	0,85	0,84	0,833
	3	4	1	0,9	0,8	0,85	0,84	0,833
	3	5	1	0,9	0,867	0,85	0,84	0,833
	4	1	1	1	0,933	0,9	0,88	0,9
	4	2	1	0,9	0,933	0,85	0,84	0,833
	4	3	1	0,9	0,867	0,85	0,84	0,833
	4	4	1	0,9	0,8	0,85	0,84	0,833
	4	5	1	0,9	0,8	0,85	0,84	0,833
	5	1	1	1	0,933	0,95	0,88	0,9
	5	2	1	0,9	0,933	0,9	0,88	0,833
	5	3	1	0,9	0,933	0,85	0,84	0,833
	5	4	1	0,9	0,867	0,85	0,84	0,833
	5	5	1	0,9	0,8	0,85	0,84	0,833
2	1	1	0,8	0,9	0,933	0,9	0,88	0,8
	1	2	0,8	0,9	0,933	0,9	0,84	0,767
	1	3	0,8	0,9	0,933	0,9	0,76	0,767
	1	4	0,8	0,9	0,933	0,9	0,76	0,767
	1	5	0,8	0,9	0,933	0,9	0,76	0,733
	2	1	1	0,9	0,933	0,95	0,88	0,833
	2	2	0,8	0,9	0,933	0,9	0,88	0,8
	2	3	0,8	0,9	0,933	0,9	0,84	0,767
	2	4	0,8	0,9	0,933	0,9	0,84	0,767
	2	5	0,8	0,9	0,933	0,9	0,8	0,767
	3	1	1	0,9	0,933	0,95	0,88	0,867
	3	2	0,8	0,9	0,933	0,95	0,88	0,8
	3	3	0,8	0,9	0,933	0,9	0,88	0,8
	3	4	0,8	0,9	0,933	0,9	0,84	0,767
	3	5	0,8	0,9	0,933	0,9	0,84	0,767
	4	1	1	0,9	0,933	0,95	0,92	0,9

Q	Boost		*Precision@k* BM25F					
	Judul	Isi	*k=5*	*k=10*	*k=15*	*k=20*	*k=25*	*k=30*
	4	2	1	0,9	0,933	0,95	0,88	0,833
	4	3	0,8	0,9	0,933	0,9	0,88	0,8
	4	4	0,8	0,9	0,933	0,9	0,88	0,8
	4	5	0,8	0,9	0,933	0,9	0,84	0,767
	5	1	1	1	0,933	0,95	0,96	0,9
	5	2	1	0,9	0,933	0,95	0,88	0,833
	5	3	0,8	0,9	0,933	0,95	0,88	0,8
	5	4	0,8	0,9	0,933	0,9	0,88	0,8
	5	5	0,8	0,9	0,933	0,9	0,88	0,8
3	1	1	1	1	0,867	0,85	0,8	0,8
	1	2	1	1	0,8	0,85	0,8	0,767
	1	3	1	1	0,8	0,85	0,8	0,767
	1	4	1	1	0,8	0,85	0,84	0,733
	1	5	1	1	0,8	0,85	0,84	0,733
	2	1	1	1	0,867	0,9	0,84	0,8
	2	2	1	1	0,867	0,85	0,8	0,8
	2	3	1	1	0,8	0,85	0,8	0,8
	2	4	1	1	0,8	0,85	0,8	0,767
	2	5	1	1	0,8	0,85	0,8	0,767
	3	1	1	1	0,933	0,9	0,88	0,8
	3	2	1	1	0,867	0,85	0,8	0,8
	3	3	1	1	0,867	0,85	0,8	0,8
	3	4	1	1	0,867	0,85	0,8	0,8
	3	5	1	1	0,8	0,85	0,8	0,8
	4	1	1	1	0,933	0,95	0,92	0,833
	4	2	1	1	0,933	0,95	0,92	0,833
	4	3	1	1	0,867	0,85	0,8	0,8
	4	4	1	1	0,867	0,85	0,8	0,8
	4	5	1	1	0,867	0,85	0,8	0,8
	5	1	1	1	0,933	0,95	0,92	0,833
	5	2	1	0,9	0,933	0,9	0,88	0,8
	5	3	1	1	0,867	0,9	0,84	0,8
	5	4	1	1	0,867	0,85	0,8	0,8
	5	5	1	1	0,867	0,85	0,8	0,8
4	1	1	1	0,9	0,733	0,7	0,72	0,667
	1	2	1	0,8	0,733	0,7	0,72	0,667
	1	3	1	0,8	0,733	0,7	0,72	0,667
	1	4	1	0,8	0,733	0,7	0,72	0,667
	1	5	1	0,8	0,733	0,7	0,68	0,667
	2	1	1	1	0,733	0,7	0,72	0,667
	2	2	1	0,9	0,733	0,7	0,72	0,667
	2	3	1	0,8	0,733	0,7	0,72	0,667
	2	4	1	0,8	0,733	0,7	0,72	0,667
	2	5	1	0,8	0,733	0,7	0,72	0,667
	3	1	1	1	0,8	0,7	0,72	0,7
	3	2	1	1	0,667	0,7	0,72	0,667
	3	3	1	0,9	0,733	0,7	0,72	0,667
	3	4	1	0,9	0,733	0,7	0,72	0,667
	3	5	1	0,8	0,733	0,7	0,72	0,667
	4	1	1	1	0,8	0,75	0,76	0,7
	4	2	1	1	0,733	0,7	0,72	0,667
	4	3	1	1	0,667	0,65	0,72	0,667
	4	4	1	0,9	0,733	0,7	0,72	0,667
	4	5	1	0,9	0,733	0,7	0,72	0,667
	5	1	1	1	0,8	0,75	0,68	0,667
	5	2	1	1	0,733	0,7	0,72	0,7
	5	3	1	1	0,733	0,7	0,72	0,667
	5	4	1	0,9	0,667	0,7	0,72	0,667
	5	5	1	0,9	0,733	0,7	0,72	0,667

4.2. Pengujian *Query* Bahasa Indonesia

Pengujian pada subbab ini dilakukan untuk menguji *query* bahasa Indonesia yang mengembalikan dokumen bahasa Inggris saja dan yang mengembalikan dokumen bahasa Indonesia dan Inggris. Sebelum mendapatkan hasil dari pengujian ini maka dilakukan pengujian *precision@k* BM25F untuk setiap *query*. Pengujian ini dilakukan dengan menggunakan 10 *query* bahasa Indonesia yang ditunjukkan pada Tabel 3.

Tabel 3. Daftar *Query* Bahasa Indonesia

No	Query
1	Rekomendasi destinasi wisata
2	Makanan khas Indonesia
3	Negara bebas Visa untuk orang Indonesia
4	Tips Liburan dan Perjalanan
5	Penginapan dan Hotel di Jakarta
6	Panduan berwisata ke Korea Selatan
7	Destinasi halal dan ramah muslim
8	Promo tiket pesawat
9	Festival Budaya Tradisional
10	Mudik libur lebaran idul fitri

Hasil terjemahan dari Tabel 3 yang diterjemahkan oleh program ditunjukkan pada Tabel 4.

Tabel 4. Daftar Terjemahan *Query* Bahasa Indonesia

No	Terjemahan Query
1	Recommended tourist destinations
2	Indonesian special food
3	Visa-free countries for Indonesia
4	Holiday Travel Tips
5	Accommodation and Hotels in Jakarta
6	Free traveled to South Korea
7	Kosher and Muslim-friendly destinations
8	Promo tickets
9	Traditional Culture Festival
10	Eid Eid holiday homecoming

Hasil pengujian *query* bahasa Indonesia yang hanya mengembalikan dokumen bahasa Inggris saja ditunjukkan pada Tabel 5.

Tabel 5. Hasil Pengujian *Query* Bahasa Indonesia pada Dokumen Bahasa Inggris

Q	P @5	P @10	P @15	P @20	P @25	P @30
1	1	0,8	0,73	0,7	0,72	0,73
2	1	0,9	0,73	0,7	0,68	0,6
3	0,8	0,8	0,73	0,75	0,72	0,63
4	1	0,9	0,87	0,85	0,8	0,77
5	1	1	1	0,9	0,88	0,87
6	1	1	0,93	0,85	0,8	0,7
7	1	1	1	0,95	0,96	0,87
8	0,8	0,6	0,53	0,6	0,52	0,43
9	1	1	0,93	0,95	0,96	0,93
10	1	1	1	0,95	0,96	0,9
Rata-rata	**0,96**	**0,9**	**0,85**	**0,82**	**0,80**	**0,74**

Pengujian selanjutnya juga akan dilakukan dengan menggunakan 10 *query* bahasa Indonesia. Hasil pengujian *query* bahasa Indonesia yang mengembalikan dokumen bahasa Indonesia dan Inggris ditunjukkan pada Tabel 6.

Tabel 6. Hasil Pengujian *Query* Bahasa Indonesia pada Dokumen Bahasa Indonesia dan Inggris

Q	P @5	P @10	P @15	P @20	P @25	P @30
1	1	1	0,93	0,95	0,88	0,9
2	1	1	0,93	0,95	0,92	0,83
3	0,8	0,7	0,6	0,5	0,56	0,63
4	1	1	1	0,95	0,96	0,93
5	1	1	1	1	0,96	0,97
6	1	1	1	1	1	1
7	1	1	1	1	1	1
8	1	1	1	1	0,96	0,9
9	1	1	0,93	0,95	0,96	0,93
10	1	0,9	0,93	0,85	0,88	0,8
Rata-rata	**0,98**	**0,96**	**0,93**	**0,92**	**0,91**	**0,89**

4.3. Pengujian *Query* Bahasa Inggris

Pengujian pada subbab ini dilakukan untuk menguji *query* bahasa Inggris yang mengembalikan dokumen bahasa Indonesia saja dan yang mengembalikan dokumen bahasa Indonesia dan Inggris. Sebelum mendapatkan hasil dari pengujian ini maka dilakukan pengujian *precision@k* BM25F untuk setiap *query*. Pengujian ini akan dilakukan dengan menggunakan 10 *query* bahasa Inggris yang ditunjukkan pada Tabel 7.

Tabel 7. Daftar *Query* Bahasa Inggris

No	Query
1	Recommendation tourist destinations
2	Indonesia Traditional Foods
3	Visa-Free Countries for Indonesian
4	Holiday Tips & Travel
5	Lodging and Hotels in Jakarta
6	Travel guide to South Korea
7	Halal and muslim friendly destinations
8	Flight tickets promo
9	Traditional Cultural Festival
10	Eid Idul Fitri Holiday

Hasil terjemahan dari Tabel 7 yang diterjemahkan oleh program ditunjukkan pada Tabel 8.

Tabel 8. Daftar Terjemahan *Query* Bahasa Inggris

No	Terjemahan Query
1	tujuan rekomendasi wisata
2	Indonesia Makanan Tradisional
3	Negara Bebas Visa bagi Indonesia
4	Tips Liburan & Travel
5	Hotel penginapan dan di Jakarta
6	panduan perjalanan ke Korea Selatan
7	Halal dan muslim tujuan ramah
8	Tiket pesawat promo
9	Festival Budaya Tradisional
10	Eid Holiday Homecoming

Hasil pengujian *query* bahasa Inggris yang hanya mengembalikan dokumen bahasa Indonesia saja ditunjukkan pada Tabel 9.

Tabel 9. Hasil Pengujian *Query* Bahasa Inggris pada Dokumen Bahasa Indonesia

Q	P @5	P @10	P @15	P @20	P @25	P @30
1	1	0,9	0,93	0,95	0,92	0,9
2	1	1	0,73	0,7	0,64	0,53
3	0,8	0,6	0,53	0,55	0,52	0,5
4	1	1	0,93	0,85	0,84	0,8
5	1	1	1	0,95	0,96	0,93
6	1	1	1	1	1	1
7	1	1	1	1	1	1

8	1	1	1	1	0,96	0,93
9	1	0,9	0,87	0,7	0,72	0,63
10	0,8	0,6	0,67	0,7	0,64	0,67
Rata-rata	**0,96**	**0,9**	**0,87**	**0,84**	**0,82**	**0,79**

Pengujian selanjutnya juga akan dilakukan dengan menggunakan 10 *query* bahasa Inggris. Hasil pengujian *query* bahasa Inggris yang mengembalikan dokumen bahasa Indonesia dan Inggris ditunjukkan pada Tabel 10.

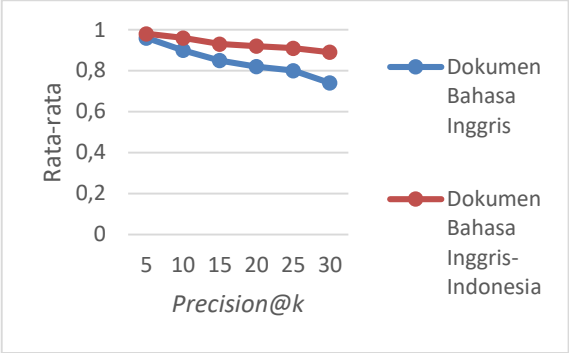
Tabel 10. Hasil Pengujian *Query* Bahasa Inggris pada Dokumen Bahasa Indonesia dan Inggris

Q	P @5	P @10	P @15	P @20	P @25	P @30
1	1	1	0,93	0,95	0,96	0,9
2	1	1	0,8	0,75	0,68	0,67
3	0,8	0,7	0,6	0,6	0,6	0,63
4	0,8	0,9	0,87	0,8	0,76	0,77
5	1	1	1	1	0,96	0,97
6	1	1	1	1	1	1
7	1	1	1	1	1	1
8	1	1	1	1	0,96	0,9
9	1	1	0,93	0,95	0,96	0,93
10	1	1	0,87	0,9	0,84	0,87
Rata-rata	**0,96**	**0,96**	**0,9**	**0,9**	**0,87**	**0,86**

4.4. Analisis

Pengujian *boost* dilakukan untuk mengetahui nilai *boost* yang paling baik untuk masing-masing *field*. *Field* yang digunakan pada penelitian ini ialah *field* judul dan isi berita. Pemilihan kedua *field* ini dikarenakan setiap berita memiliki bagian judul dan isi. Dari hasil yang didapatkan pada Tabel 2 menggunakan 4 *query* tersebut dapat diketahui bahwa hasil *precision@k* BM25F terbaik ditunjukkan saat *boost* judul bernilai 5 dan *boost* isi bernilai 1. Hal ini menandakan bahwa *field* judul lebih berperan penting pada penelitian ini. Judul berita dianggap sudah dapat mencerminkan keseluruhan isi berita. Nilai *boost* ini digunakan untuk pengujian query bahasa Indonesia dan bahasa Inggris.

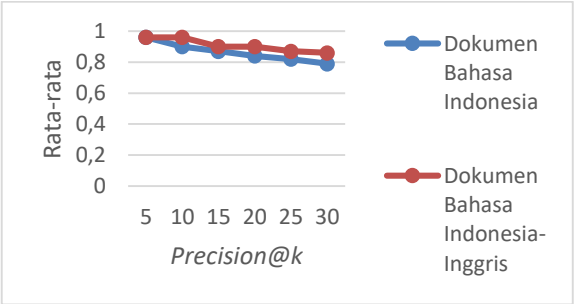
Pengujian *precision@k* BM25F pada *query* bahasa Indonesia yang mengembalikan dokumen bahasa Inggris didapatkan nilai rata-rata secara berurutan yaitu 0,96, 0,9, 0,85, 0,82, 0,8, serta 0,74 dan pada *query* bahasa Indonesia yang mengembalikan dokumen bahasa Indonesia dan Inggris didapatkan nilai rata-rata dengan urutan 0,98, 0,96, 0,93, 0,92, 0,91, dan 0,89. Grafik pengujian *query* bahasa Indonesia ditunjukkan pada Gambar 3.



Gambar 3. Grafik Perbandingan Pengujian *Query* Bahasa Indonesia

Dari kedua hasil tersebut diketahui bahwa hasil pengujian *query* bahasa Indonesia pada dokumen berita bahasa Indonesia dan Inggris lebih baik dibandingkan hanya dengan menggunakan dokumen bahasa Inggris saja dikarenakan jumlah data relevan yang lebih banyak. Tetapi, dapat diketahui pula bahwa hanya dengan menggunakan dokumen berita bahasa Inggris saja sistem temu kembali informasi lintas bahasa yang dibuat sudah mampu mengembalikan dokumen-dokumen yang relevan dengan cukup baik. Rata-rata nilai *precision@k* tertinggi sebesar 0,98 didapatkan saat *k*=5. Hal ini disebabkan karena data relevan tersebar pada 5 dokumen teratas.

Pengujian *precision@k* BM25F pada *query* bahasa Inggris yang mengembalikan dokumen bahasa Indonesia didapatkan nilai rata-rata secara berurutan yaitu 0,96, 0,9, 0,87, 0,84, 0,82, serta 0,79 dan pada *query* bahasa Inggris yang mengembalikan dokumen bahasa Indonesia dan Inggris didapatkan nilai rata-rata secara berurutan yaitu 0,96, 0,96, 0,9, 0,9, 0,87, dan 0,86. Grafik pengujian *query* bahasa Inggris ditunjukkan pada Gambar 4.

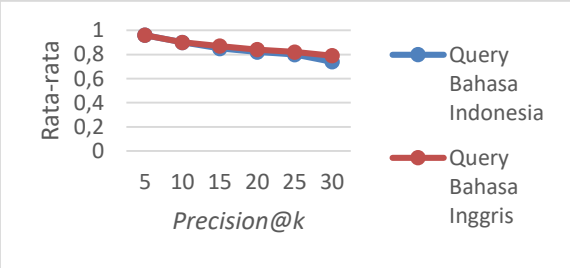


Gambar 4. Grafik Perbandingan Pengujian *Query* Bahasa Inggris

Dari kedua hasil tersebut diketahui bahwa hasil pengujian *query* bahasa Inggris pada dokumen berita bahasa Indonesia dan Inggris

lebih baik dibandingkan hanya dengan menggunakan dokumen bahasa Indonesia saja. Tetapi, dapat diketahui pula bahwa hanya dengan menggunakan dokumen berita bahasa Indonesia saja sistem temu kembali informasi lintas bahasa yang dibuat sudah mampu mengembalikan dokumen-dokumen yang relevan dengan cukup baik. Rata-rata nilai *precision@k* tertinggi sebesar 0,96 didapatkan saat *k*=5. Hal ini disebabkan karena data relevan tersebar pada 5 dokumen teratas.

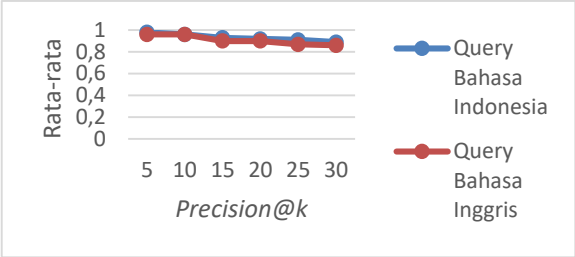
Hasil perbandingan nilai rata-rata *query* bahasa Indonesia yang hanya mengembalikan dokumen bahasa Inggris dan nilai rata-rata *query* bahasa Inggris yang hanya mengembalikan dokumen bahasa Indonesia ditunjukkan pada Gambar 5.



Gambar 5. Grafik Perbandingan Nilai Rata-rata *Query* dengan Dokumen Satu Bahasa

Diketahui bahwa *query* bahasa Inggris yang mengembalikan dokumen bahasa Indonesia mengembalikan hasil yang lebih baik. Hal ini disebabkan karena penerjemahan dari bahasa Inggris ke bahasa Indonesia lebih baik sehingga dokumen relevan yang mengandung kata-kata tersebut lebih banyak dikembalikan. Contohnya seperti *query* "*Lodging and Hotels in Jakarta*" yang diterjemahkan menjadi "Hotel penginapan dan di Jakarta", "*Halal and muslim friendly destinations*" yang diterjemahkan menjadi "Halal dan muslim tujuan ramah", dan "*Travel guide to South Korea*" yang diterjemahkan menjadi "panduan perjalanan ke Korea Selatan". Hasil terjemahan tersebut memang tidak sesuai tata letak struktur bahasa Indonesia, namun sudah cukup akurat karena mengandung kata-kata yang sesuai.

Hasil perbandingan nilai rata-rata *query* bahasa Indonesia dan *query* bahasa Inggris yang mengembalikan dokumen berita bahasa Indonesia dan Inggris ditunjukkan pada Gambar 6.



Gambar 6. Grafik Perbandingan Nilai Rata-rata *Query* dengan Dokumen Dua Bahasa

Diketahui bahwa *query* bahasa Indonesia yang mengembalikan dokumen bahasa Indonesia dan Inggris mengembalikan hasil yang lebih relevan dibandingkan *query* bahasa Inggris. Hal ini disebabkan karena lebih banyaknya dokumen bahasa Indonesia relevan yang dikembalikan. Contohnya seperti pada *query* "Makanan khas Indonesia" yang mengembalikan 19 dokumen bahasa Indonesia dan terdapat 17 dokumen yang relevan. Contoh lainnya yaitu pada *query* "Tips Liburan dan Perjalanan" yang mengembalikan 17 dokumen bahasa Indonesia dan terdapat 17 dokumen yang relevan.

5. KESIMPULAN DAN SARAN

Kesimpulan yang didapatkan dari penelitian ini ialah sistem temu kembali informasi lintas bahasa yang telah dibuat dengan menggunakan metode BM25F terbukti sudah mampu mengembalikan dokumen yang relevan. Dokumen didapatkan hanya dengan memasukkan *query* dalam satu bahasa kemudian sistem secara otomatis akan mengembalikan dokumen yang relevan. Nilai *boost* yang paling optimal ialah saat nilai *boost field* judul bernilai 5 dan *field* isi bernilai 1.

Pengujian yang mengembalikan dokumen dua bahasa menghasilkan nilai rata-rata *precision* yang lebih tinggi dibandingkan dengan yang mengembalikan dokumen dalam satu bahasa saja. Hal ini disebabkan lebih banyaknya dokumen berita relevan yang dikembalikan. Nilai rata-rata *precision@k* yang paling tinggi dihasilkan pada pengujian *query* bahasa Indonesia yang mengembalikan dokumen bahasa Indonesia dan Inggris. Nilai rata-rata yang dihasilkan secara berurutan ialah 0,98, 0,96, 0,93, 0,92, 0,91, dan 0,89. Saat nilai *k*=5 didapatkan hasil tertinggi yaitu 0,98. Hal ini disebabkan karena data relevan tersebar pada 5 dokumen teratas. Saran untuk penelitian selanjutnya yaitu menggunakan rumus IDF Modifikasi untuk menghindari nilai negatif pada

metode BM25F yang diharapkan dapat memberikan hasil yang lebih baik walaupun dengan data dan topik yang berbeda.

6. DAFTAR PUSTAKA

CLEF, 2003. *Cross Language Evaluation Forum.* [Online] Tersedia di: <http://www.clef-campaign.org/2003.htm> [Diakses 25 Agustus 2019].

Elayeb, B. & Bounhas, I., 2015. Arabic Cross-Language Information Retrieval: A Review. *ACM Transaction on Asian and Low-Resource Language Information Processing,* 15(3), pp. 18-44.

Fang, H., Tao, T. & Zhen, C., 2004. *A Formal Study of Information Retrieval Heuristics.* New York, Association for Computing Machinery, pp. 49-56.

Fesenmaier, D., Xiang, Z., Pan, B. & R., L., 2010. A Framework of Search Engine Use for Travel Planning. *Journal of Travel Research,* 50(6), p. 587–601.

Ganesan, K., 2019. *All you need to know about text preprocessing for NLP and Machine Learning.* [Online] Tersedia di: <https://www.kdnuggets.com/2019/04/text-preprocessing-nlp-machine-learning.html> [Diakses 18 November 2019].

Garcia, E., 2011. *A Tutorial on the BM25F Model.* [Online] Tersedia di: <http://www.minerazzi.com/tutorials/bm25f-model-tutorial.pdf> [Diakses 9 Oktober 2019].

Iriani, P. R., Indriati & Adikara, P. P., 2019. Temu Kembali Informasi Lintas Bahasa untuk Dokumen Berita Berbahasa Indonesia-Inggris Menggunakan Metode BM25. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer,* 3(5), pp. 4383-4390.

Itakura, K. Y. & Clarke, C. L., 2010. *A framework for BM25F-based XML retrieval.* Geneva, SIGIR 2010, pp. 843-844.

KBBI, 2019. *Arti kata berita - Kamus Besar Bahasa Indonesia (KBBI) Online.* [Online] Tersedia di: <https://kbbi.web.id/berita> [Diakses 9 Oktober 2019].

Manning, C. D., Raghavan, P. & Schütze, H., 2009. *Introduction to information retrieval.* Online ed. Cambridge: Cambridge University Press.

Nie, J.-Y., 2010. *Cross-Language Information Retrieval.* Canada: Morgan & Claypool.

Pérez-Agüera, J. R. et al., 2010. *Using BM25F for semantic search.* New York, ACM Press, pp. 1-8.

Pérez-Iglesias, J., Pérez-Agüera, J. R., Fresno, V. & Feinstein, Y. Z., 2009. *Integrating the Probabilistic Models BM25/BM25F into Lucene.* [Online] Tersedia di: <https://arxiv.org/abs/0911.5046> [Diakses 9 Oktober 2019].

Robertson, S. & Zaragoza, H., 2009. The Probabilistic Relevance Framework:BM25 and Beyond. *Foundations and Trends in Information Retrieval,* 3(4), pp. 333-389.

Sari, S. & Adriani, M., 2014. Learning to Rank for Determining Relevant Document in Indonesian-English Cross Language Information Retrieval using BM25. *IEEE,* pp. 309-314.

Wahyudi, J., 1991. *Komunikasi Jurnalistik Pengetahuan Praktis Kewartawanan Surat Kabar, Majalah, Radio, dan Televisi.* Bandung: Alumni.

Wang, J. & Oard, D. W., 2012. Matching meaning for cross-language information retrieval. *Information Processing & Management,* 48(4), pp. 631-653.

Zhou, D. et al., 2012. Translation Techniques in Cross-Language Information Retrieval. *ACM Computing Surveys (CSUR),* 45(1), pp. 1-44.