Bab 5

Penutup

5.1 Kesimpulan

Pada penelitian ini, dengan menggunakan berbagai metode atau teknik perolehan informasi, yaitu operator kedekatan kata, umpan balik relevan semu, pendeteksian bahasa, analisis pranala, *PageRank*, dan pengelompokan dokumen, diperoleh hasil evaluasi yang sangat bervariasi, apakah itu ketika dilakukan evaluasi secara umum ataupun berdasarkan setiap bahasa. Penelitian ini juga menghasilkan kesimpulan bahwa Indri *Search Engine* dapat memperoleh dokumen multibahasa dan teknik perolehan informasi Web standar dapat diterapkan pada koleksi dokumen Web multibahasa.

Dari eksperimen yang dilakukan, secara umum dapat diambil kesimpulan bahwa dengan mengkombinasikan teknik perolehan informasi operator kedekatan kata `#odN` dan `#uwN` dengan pengurutan ulang menggunakan pendeteksian bahasa dapat meningkatkan nilai perolehan informasi pada koleksi dokumen Web multilingual sebesar 0.72% dengan menggunakan kombinasi dengan operator kedekatan kata `#odN`. Sedangkan teknik pengurutan ulang seperti analisis pranala, *PageRank*, dan pengelompokan dokumen tidak dapat menghasilkan penurunan pada nilai perolehan informasi sebesar -46.3431% - -0.60%. Untuk keseluruhan hasil evaluasi secara umum terhadap seluruh teknik dapat dilihat pada bagian 4.8 pada tabel 4.20.

Dengan menggunakan teknik umpan balik relevan semu, secara umum nilai perolehan informasi dapat ditingkatkan. Akan tetapi peningkatan itu terjadi apabila kita mendefinisikan bobot antara kueri awal dan kueri perluasan yang terbentuk setelah umpan balik. Bobot kueri awal yang ideal untuk mendapatkan kenaikan pada nilai perolehan informasi

adalah 0.9. Selain pengaturan bobot, jumlah dokumen dan kata pada peringkat teratas perlu diperhatikan juga. Dari eksperimen yang dilakukan peningkatan terjadi pada saat menggunakan 20 dokumen dan 5 s/d 10 kata pada peringkat teratas sebesar 1.02% - 1.61%. Apabila jumlah kata pada peringkat teratas dibesarkan, maka nilai perolehan informasi akan semakin menurun.

Dengan melakukan evaluasi berdasarkan bahasa, penggunaan setiap teknik memiliki keunggulan dibandingkan teknik lainnya berdasarkan setiap bahasa. Data mengenai evaluasi setiap teknik berdasarkan bahasa dapat dilihat pada tabel 4.21 dan 4.22. Berikut adalah kesimpulan mengenai efek penggunaan setiap teknik terhadap perolehan informasi jika dievaluasi berdasarkan bahasa:

- Dengan mengkombinasikan teknik operator kedekatan kata dan teknik pengurutan ulang pendeteksian bahasa, peningkatan nilai perolehan informasi terdapat pada 6 bahasa sebesar 0.1478% - 27.1431%, dan penurunan terdapat pada 10 bahasa sebesar -0.7066% - -0.0156%.

- Dengan mengkombinasikan teknik operator kedekatan kata dan teknik pengurutan ulang analisis pranala dalam, peningkatan nilai perolehan informasi terdapat pada 4 bahasa sebesar 6.5585% - 77.1804%, dan penurunan terdapat pada 4 bahasa sebesar -60.9202% - -0.3339%.

- Dengan mengkombinasikan teknik operator kedekatan kata dan menggunakan teknik pengurutan ulang analisis pranala luar, peningkatan nilai perolehan informasi terdapat pada 9 bahasa sebesar 0.7151% - 134.9881%, dan penurunan terdapat pada 6 bahasa sebesar -23.2941% - -1.6330%.

- Dengan mengkombinasikan teknik operator kedekatan kata dan gabungan teknik analisis pranala dalam dan luar, peningkatan nilai perolehan informasi terdapat pada 5 bahasa sebesar 0.2809% - 100%, dan penurunan terdapat pada 10 bahasa sebesar -92.0830% - -3.1030%.

- Dengan mengkombinasikan teknik operator kedekatan kata dan teknik pengurutan ulang *PageRank*, peningkatan nilai informasi terdapat pada 4 bahasa sebesar 5.1083% - 77.1804%,dan penurunan terdapat pada 10 bahasa sebesar -96.4844% - -2.6968%.

- Dengan mengkombinasikan teknik operator kedekatan kata dan teknik pengurutan ulang pengelompokan berdasarkan bahasa, peningkatan nilai informasi terdapat pada 3 bahasa sebesar 3.6440% - 38.7337%, dan penurunan terdapat pada 9 bahasa sebesar -53.6322% - -1.6715%.

- Dengan menggunakan teknik umpan balik relevan semu yang menghasilkan nilai paling optimal yaitu dengan menggunakan 20 dokumen dan 5 kata teratas serta bobot kueri awal 0.9, peningkatan nilai perolehan informasi terdapat pada 7 bahasa sebesar 0.0472% - 5.1523%, penurunan hanya terdapat pada dua bahasa sebesar -4.2597% - -0.2760%.

Pada eksperimen ini, dalam penggunaan teknik pengurutan ulang dengan pengelompokan dokumen perolehan berdasarkan topik, penulis tidak bisa menampilkan hasil yang dapat merepresentasikan semua bahasa yang terdapat pada koleksi. Dikarenakan koleksi yang digunakan pada eksperimen teknik ini adalah sampel data. Dari sampel data yang dihasilkan terdapat 7 bahasa yang terdeteksi, yaitu Denmark, Belanda, Inggris, Jerman, Portugis, Rusia, dan Spanyol. Peningkatan nilai perolehan informasi hanya terdapat pada bahasa satu bahasa, yaitu Spanyol sebesar 12.4754% dan penurunan terdapat pada 5 bahasa sebesar -97.5600% - -35.6133% (dapat dilihat pada tabel 4.15), sedangkan pada bahasa lain mengalami penurunan nilai perolehan informasi.

Untuk bahasa Yunani, dengan meggunakan teknik-teknik perolehan informasi dalam penelitian ini, tetap tidak dapat diperoleh dokumen yang relevan sama sekali. Hal ini menunjukkan bahwa untuk bahasa Yunani, Indri *Search Engine* yang digunakan tidak dapat memperoleh dokumen Yunani yang relevan dengan kueri Yunani yang diberikan pencari informasi.

5.2 Saran dan Kendala

Masalah yang menjadi isu utama dalam penelitian ini adalah proses pendeteksian bahasa terhadap koleksi dokumen yang digunakan. Proses pendeteksian bahasa memegang peranan yang sangat penting dalam perolehan informasi Web multilingual, dengan memperbaiki proses ini nilai perolehan informasi yang menggunakan teknik-teknik pengurutan ulang seperti pendeteksian bahasa dan pengelompokan bahasa akan dapat ditingkatkan. Dengan mela-

kukan pelatihan terhadap program pendeteksian bahasa aplikasi TextCat, nilai informasi akan dapat ditingkatkan lagi (lihat bagian 4.3).

Untuk teknik pengurutan ulang yang menggunakan pranala, seperti analisis pranala dalam, analisis pranala luar, dan *PageRank*, peningkatan nilai informasi dapat ditingkatkan dengan memperbaiki proses awal dalam mengambil pranala dalam dan luar dari koleksi dokumen. Pada penelitian ini penulis menggunakan aplikasi `harvestlinks` untuk mengambil semua pranala dari koleksi dokumen, penulis menyarankan untuk memperbaiki aplikasi yang digunakan atau dengan menggunakan aplikasi lain yang memiliki fungsi yang sama, agar perolehan pranala dari suatu koleksi dokumen jauh lebih baik lagi (lihat bagian 4.4).

Teknik pengelompokan berdasarkan topik dapat ditingkatkan dengan menggunakan algoritma pengelompokan dokumen yang lebih baik lagi daripada yang digunakan pada penelitian ini yaitu *centroid*. Selain itu dalam melakukan perbandingan kemiripan kueri dan topik pada setiap *centroid*, diperlukan pemilihan kata kunci yang tepat pada setiap *centroid* (lihat bagian 4.6.1).

Berbagai hasil penelitian yang dilakukan dengan menggunakan lima teknik tersebut diharapkan dapat dikembangkan lebih lanjut pada penelitian selanjutnya. Penelitian dapat dikembangkan dengan mengkombinasikan kombinasi-kombinasi yang terbaik dari masing-masing teknik, yaitu dengan mengkombinasikan operator kedekatan kata, penggunaan struktur dokumen HTML, penggunaan umpan balik relevan semu, dan teknik pendeteksian bahasa untuk melakukan pengurutan ulang. Selain itu, dengan menemukan pemecahan masalah terhadap bahasa-bahasa yang memiliki penulisan non-latin seperti Yunani dan *Russian* dalam perolehan informasi Web dapat dilakukan pada penelitian selanjutnya.

Selama penelitian ini berlangsung, penulis menggunakan server GRID[1] Universitas Indonesia bernama Hastinapura. Penggunaan server ini dikarenakan penulis membutuhkan sumber daya yang besar dalam mengolah data yang besar. Selama penelitian berlangsung penulis mengalami kendala pada waktu proses yang lama dan ruang penyimpanan yang terbatas pada server Hastinapura. Masalah pada waktu proses diakibatkan karena sumber daya yang ada pada server Hastinapura dirasakan kurang untuk memproses data yang cukup besar. Ruang penyimpanan pada server dirasakan kurang mencukupi karena penulis menggunakan koleksi dokumen yang cukup besar, dan setiap proses pada koleksi dokumen tersebut juga akan menghabiskan ruang penyimpanan. Pada penelitian ini, penulis harus menghapus

[1]Lihat "http://grid.ui.ac.id/".

beberapa hasil eksperimen yang sudah tercatat untuk meluangkan ruang penyimpanan server Hastinapura. Akan tetapi, dengan menggunakan server GRID Hastinapura eksperimen dapat dijalankan jauh lebih cepat daripada menggunakan komputer biasa, ruang penyimpanan yang cukup besar dapat memfasilitasi penulis dalam menyimpan koleksi dokumen maupun data-data hasil eksperimen yang belum sempat tercatat.