BAB 1
PENDAHULUAN

1.1 Latar Belakang

Jumlah pengguna internet di dunia semakin lama semakin meningkat drastis, khusus untuk negara Indonesia, pada akhir 2013 pengguna internet telah mencapai 71,19 juta (Pitoyo 2014). Kondisi tersebut mendorong pertumbuhan yang besar pada artikel-artikel *web*. *Web* merupakan gudang artikel yang berisi ribuan atau bahkan jutaan artikel. Banyaknya artikel yang ada di web melahirkan kebutuhan baru bagi pengguna. Salah satu kebutuhan baru yang muncul adalah informasi mengenai artikel terkait yang sedang kita buka, tentunya pengguna artikel akan merasa nyaman ketika artikel yang mereka baca juga memberikan informasi mengenai artikel sejenis yang berada dalam satu topik. Kebutuhan lainnya adalah pada proses pencarian artikel. Seorang pencari artikel terkadang tidak secara tepat memasukkan kata kunci pada mesin pencari, mengingat setiap orang memiliki banyak pilihan kata dalam mengungkapkan ide mereka. Jika sebuah mesin pencari artikel mampu memberikan hasil pencarian yang lebih kaya, dimana hasil pencarian tidak harus sama persis dengan kata kunci, namun masih memiliki kedekatan topik dengan kata kunci, maka hal ini juga dapat membantu pencari untuk menemukan artikel yang menurut mereka relevan.

Agar dapat memenuhi kebutuhan-kebutuhan di atas, maka diperlukan sebuah proses pengelompokan artikel sesuai dengan topiknya. Akan tetapi timbul masalah lain ketika jumlah artikel yang harus dikelompokkan jumlahnya sangat banyak sementara waktu yang tersedia terbatas. Kemampuan manusia memiliki banyak keterbatasan untuk melakukannya, sehingga diperlukan bantuan mesin komputer untuk menyelesaikan pekerjaan tersebut. Proses pengelompokan artikel merupakan proses yang sangat kompleks, sehingga penggunaan logika *if* saja tidak akan mampu menyelesaikannya, oleh sebab itu komputer perlu diberikan kecerdasan buatan agar dapat menyelesaikan proses kompleks tersebut.

Beberapa upaya telah dilakukan untuk dapat mengelompokkan artikel berbahasa Indonesia secara otomatis. (Arifin dan Setiono 2002) mencoba untuk mengelompokkan artikel kejadian (*event*) dengan menggunakan *single pass clustering*, upaya yang dilakukan adalah mengekstraksi ciri-ciri artikel berdasarkan

kemunculan kata. Upaya tersebut mampu mengembalikan dokumen pada kelompoknya (*recall*) sebesar 76% dan memiliki ketepatan (presisi) sebesar 87%. (Liliana, Hardianto dan Ridok 2011) dalam penelitiannya menggunakan metode *Support Vector Machine* untuk mengelompokkan artikel berdasarkan kemunculan kata dan diperoleh akurasi sebesar 91,67 %, namun jenis pengelompokan yang dilakukan adalah jenis pengelompokan yang menggunakan pembelajaran tersupervisi. *Self Organizing Map* (*SOM*) merupakan salah satu pendekatan dalam pembelajaran tanpa supervisi dengan cara memetakan data berdimensi banyak menjadi lebih sederhana. Pendekatan ini telah diterapkan oleh (Ambarwti dan Winarko 2014) untuk dapat mengelompokkan artikel berbahasa Indonesia berdasarkan kemunculan kata yang terkandung pada seluruh dokumen.

Seluruh upaya di atas menggunakan kemunculan kata sebagai parameter ciri, selain menggunakan kemunculan kata sebagai dasar pengelompokan artikel, ciri lain yang bisa digunakan adalah berdasarkan struktur urutan kata. Mendeteksi urutan kata dapat dilakukan dengan merepresentasikan struktur urutan ke dalam struktur data berbentuk pohon (*Suffix Tree*), maupun indeks urutan kata (*Document Indeks Graph*). *Suffix Tree Clustering* (Oren, Oren dan Omid, et al. 1997), *Document Index Graph* (Khaled M. Hammouda 2004) dan matriks *suffix tree* (Chim dan Deng 2008) merupakan salah satu upaya pengelompokan artikel yang menggunakan urut-urutan kata, namun ciri tersebut tidak dapat berdiri sendiri sebagai parameter. Ciri tersebut harus tetap dikombinasikan dengan kemunculan kata.

Dari serangkaian upaya yang telah dilakukan dapat disimpulkan bahwa kemunculan kata masih sangat dominan untuk digunakan sebagai ciri dalam pengelompokan dokumen. Meskipun demikian, kemunculan kata tidak selalu menjadi ukuran dalam menentukan relevansi artikel terkait dan penentuan hasil pencarian. Pada kasus penentuan artikel terkait, dua artikel dengan topik yang sama tidak selalu memiliki kesamaan kemunculan kata. Begitu juga pada kasus pencarian, dimana hasil pencarian yang relevan bagi pengguna terkadang sama sekali tidak mengandung kata kunci pencarian, namun memiliki kedekatan topik dengan kata kunci pencarian. Kesamaan maupun perbedaan konteks kalimat yang digunakan oleh kata kunci dengan kata lainnya menyebabkan kedekatan topik

tersebut muncul, contoh : kata mobil sering digunakan pada konteks yang sama dengan kata kendaraan, atau mungkin sebaliknya, kata bank terkadang digunakan pada beberapa konteks yang berbeda seperti bank soal dan bank instansi keuangan. Kesamaan maupun perbedaan penggunaan konteks salah satunya dapat dilihat dari latensi. Latensi dapat dipandang sebagai hubungan yang muncul antara dua buah artikel yang tidak memiliki kesamaan kemunculan kata, dimana hubungan tersebut muncul karena terdapat artikel ke tiga yang memiliki kesamaan kemunculan kata dengan artikel ke satu dan ke dua.

Dari sini dapat disimpulkan bahwa latensi berpotensi untuk dijadikan sebagai dasar dalam pengelompokan dokumen. Setelah diketahui ciri berdasarkan latensi hal yang tidak kalah pentingnya adalah proses pengelompokan itu sendiri. (Bakhshi, Feizi-Derakhshi dan Zafarani 2012) melalui pengujian yang telah dilakukan menyimpulkan bahwa SOM mampu melakukan kluster dengan akurasi yang lebih baik dibanding *K-Means*, *Single-Linkage*, maupun DBSCAN. Pada percobaan lain, (Abbas 2008) juga menyimpulkan bahwa SOM memiliki sensitivitas terhadap *noise* yang lebih rendah jika dibandingkan dengan kluster secara hirarki, pendekatan ini juga sesuai jika digunakan pada data berukuran besar. Melalui pengelompokan berdasarkan latensi diharapkan kebutuhan-kebutuhan terkait penelusuran dan penentuan artikel terkait dapat terpenuhi.

1.2 Rumusan Masalah

Berdasarkan latar belakang permasalahan yang ada dalam penelitian ini dapat dirumuskan sebuah permasalahan yaitu :

1. Bagaimana memunculkan ciri latensi yang terkandung pada kumpulan artikel, selanjutnya dilakukan kluster dengan pendekatan *SOM*.
2. Bagaimana kualitas pengelompokan artikel ketika sebelum dengan sesudah memunculkan ciri latensi.

1.3 Tujuan Penelitian

Tujuan yang ingin dicapai pada penelitian ini adalah membangun sebuah program komputer :

1. Untuk mendeteksi kemiripan antar artikel bukan hanya dari sudut pandang kemunculan kata saja, tetapi juga dari latensi.

2. Untuk mengevaluasi tingkat *recall* dan presisi antara sebelum dan sesudah munculnya latensi antar artikel.

1.4 Manfaat Penelitian

Manfaat yang ingin dicapai dari penelitian ini adalah :

1. Pengelompokan yang didasarkan pada kedekatan kemunculan kata dapat dikembangkan menjadi sebuah aplikasi pengelola berita maupun artikel.
2. Pola yang terbentuk dari proses memunculkan hubungan tersembunyi dapat dikembangkan untuk penelitian lebih lanjut.

1.5 Batasan Masalah

Batasan masalah pada penelitian ini adalah :

1. Artikel yang akan dikelompokkan memiliki struktur bahasa yang baku.
2. Sumber artikel diambil dari media massa elektronik detik, metrotvnews dan liputan6.
3. Proses pengelompokan tidak melibatkan kata penghubung, sehingga menggunakan kamus kata penghubung untuk menyaring kata.
4. Proses pengelompokan tidak dilakukan secara *real time*, sehingga data artikel terlebih dahulu disimpan pada *database*.