

BAB II

TINJAUAN PUSTAKA

Pada bab ini menjelaskan topik taksonomi yang merupakan pengorganisasian informasi yang penting karena merupakan dasar dalam memahami suatu informasi. Taksonomi membantu memahami isi berdasarkan hubungan hierarkis serta hubungan yang bersilangan. Rekayasa perangkat lunak sendiri memiliki beberapa taksonomi yaitu SWEBOOK, *Computer Curriculum - Software Engineering* (SE2004) dan *Computing Classification System* versi 1998 (CCS98) yang berlaku sampai tahun 2011 (R. & Mardiyanto, 2011).

Taksonomi yang dibuat oleh pengguna dimanfaatkan untuk mengkategorikan dan mengambil isi web seperti halaman Web dan link Web, menggunakan label terbuka yang disebut tag. Biasanya, taksonomi yang berbasis internet, tetapi mereka mungkin digunakan dalam konteks lain. Tujuan *taksonomy tagging* adalah membuat informasi semakin mudah untuk mencari, menemukan dan navigasi dari waktu ke waktu. Sebuah berkembang dengan baik taksonomy idealnya diakses sebagai bersama kosakata yang baik berasal oleh, dan akrab bagi, yang pengguna utama (Wikipedia). Contoh lain kolaborasi terbuka, terdengar seperti anarki di Dunia Wide Web, dengan orang-orang membungkuk aturan untuk masing-masing kebutuhan dan selera, tetapi tidak. Bahkan, penandaan adalah pada intinya dari beberapa komunitas online yang paling bersemangat dan kohesif. Tentu saja itu terletak jauh dari kenyamanan terstruktur dari taksonomi konvensional dikontrol, seperti yang pustakawan tetapi

berguna untuk membuat digital berselancar dikelola (Dye, 2006).

Menurut (Raharjo & Winarko, 2014) klasifikasi adalah proses menentukan suatu obyek kedalam suatu kelas atau kategori yang telah ditentukan. Penentuan obyek dapat menggunakan suatu model tertentu bebeapa model yang bisa digunakan antara lain: classification (IF-THEN) rules, dan decision trees. Klasifikasi data atau dokumen dimulai dengan membangun aturan klasifikasi dengan algoritma klasifikasi tertentu menggunakan data training (tahapan ini sering disebut dengan tahapan pembelajaran) dan tahap pengujian algoritma dengan data testing.

Menurut (Setiawan, 2015) dalam penelitiannya aplikasi yang dibangun memanfaatkan teknologi web scraping untuk mengambil data pada berbagai situs berita. Data berita yang diambil tersebut langsung masuk ke dalam basis data untuk disimpan. Data berita dalam basis data dapat ditampilkan pada aplikasi. Selain ditampilkan, data berita dapat diolah untuk di ringkas menjadi lebih mudah mengetahui inti yang dicari.

Dimulai dengan memilih sejumlah variabel yang menutupi dimensi dan yang akan membantu untuk mengembangkan taksonomi tersebut. Beberapa dari mereka memiliki fitur umum untuk tiga konteks pendekatan yaitu, (nilai-nilai, belajar, difusi), sementara yang lain menyajikan perbedaan yang relevan yaitu otoritas, kepemilikan, kekayaan intelektual, kolaborasi, dan demokratisasi. Menurut evolusi alat-alat dan kolaborasi kegiatan, (Bernard, et al., 2005) mengklasifikasikan

dua dimensi ke dalam konektivitas sosial dan potensi berbagi informasi atau informasi konektivitas. Dengan demikian, kita memiliki tiga kuadran yaitu daerah konteks, kaya informasi, dan pengetahuan atau interaksi sosial yang terlokasi.

Menurut (Kurniawan, et al., 2012) dalam penelitiannya yang berjudul klasifikasi konten berita dengan metode *text mining* menjelaskan pengklasifikasian dapat diperoleh menggunakan metode *text mining* yang merupakan variasi dari data mining yang berusaha menemukan pola yang menarik dari sekumpulan data tekstual yang berjumlah besar [3]. Langkah-langkah yang dilakukan dalam *text mining* adalah sebagai berikut

1. *Text Preprocessing*

Tindakan yang dilakukan pada tahap ini adalah *toLowerCase*, yaitu mengubah semua karakter huruf menjadi huruf kecil, dan *Tokenizing* yaitu proses penguraian deskripsi yang semula berupa kalimat - kalimat menjadi kata-kata dan menghilangkan delimiter-delimiter seperti tanda titik(.), koma(,), spasi dan karakter angka yang ada pada kata tersebut [7].

2. *Feature Selection*

Pada tahap ini tindakan yang dilakukan adalah menghilangkan *stopword* (*stopword removal*) dan *stemming* terhadap kata yang berimbuhan [3][4]. *Stopword* adalah kosakata yang bukan merupakan ciri (kata unik) dari suatu dokumen [5]. Misalnya "di", "oleh", "pada", "sebuah", "karena" dan lain sebagainya.

Intelijen berbagi lokasi (*metaweb*) berarti masa depan dan evolusi kolaboratif paradigma. Pendekatan serupa diikuti oleh (Befautsou & Mentzas, 2002), yang mengklasifikasikan alat kolaborasi oleh fokus mereka pada kolaborasi dan informasi manajemen. Tapi bagaimana kita bisa berhubungan analisis kami untuk evolusi ini? Kami telah mencoba untuk menganalisis bagaimana dimensi ini berhubungan dengan variable yang menunjukkan bagaimana variabel-variabel terkait dengan kedua dimensi.

Di antara berbagai metode pembelajaran tanpa pengawasan, teknik *clustering* dari tertentu popularitas untuk pendekatan dokumen-berpusat. (Chirita, et al., 2007), menyarankan metode bernama P-TAG untuk secara otomatis menghasilkan personalisasi tag secara semantik. Mereka membayar perhatian khusus untuk personalisasi anotasi dari halaman web. Dalam pendekatan berorientasi dokumen mereka, halaman web adalah dibandingkan dengan dokumen desktop yang baik menggunakan kesamaan kosinus atau laten semantik analisa. Kata kunci kemudian diekstraksi dari dokumen yang sama untuk rekomendasi. Pendekatan kata kunci berorientasi kedua alternatif menemukan terjadinya istilah dalam dokumen dan berbeda merekomendasikan tag tersisa dari sejenis dokumen desktop ke halaman web. Pendekatan *hybrid* ketiga menggabungkan sebelumnya dua metode. Dari titik penyaringan kolaboratif pandang, dua yang pertama Metode dapat diartikan sebagai CF berbasis-item dengan item yang dokumen dan kata kunci masing-masing. metode mereka, bagaimanapun, tidak menyelidiki perilaku antara pengguna yang berbeda untuk

halaman web yang sama. Pendekatan berbasis pengelompokan diusulkan dalam (Begelman, et al., 2006) untuk agregat semantik terkait tag pengguna ke kelompok yang sama. *Tags* direpresentasikan sebagai grafik di mana setiap node adalah tag dan tepi antara dua node sesuai dengan mereka co-kejadian dalam dokumen yang sama. *Tags* di *cluster* yang sama yang direkomendasikan ke pengguna berdasarkan kesamaan mereka. Demikian pula, metode penjelasan otomatis untuk gambar diusulkan pada (Li & Wang, 2006). Sebuah model generatif dilatih oleh mengeksploitasi hubungan statistik antara kata-kata dan gambar. Sebuah distribusi diskrit (D2) algoritma clustering diperkenalkan untuk prototipe berbasis cluster gambar dan kata-kata, sehingga model yang sangat efisien untuk tag gambar.

Pada tabel 2.1, dapat dilihat perbandingan penelitian tentang pembangunan aplikasi pemrosesan data sebelumnya dengan yang akan dibuat oleh penulis

Tabel 2.1 Perbandingan Penelitian

no	Perbandingan	(Kurniawan, et al., 2012)	(Raharjo & Winarko, 2014)	(Setiawan, 2015)	Penulis (2016)
1.	Judul	Klasifikasi Konten Berita Dengan Metode Text Mining	Klasterasi, Klasifikasi Dan Peringkat Teks Berbahasa Indonesia	Pembangunan Perangkat Lunak Web Scraping Untuk Situs Berita dan Peringkat Berita	Pembangunan Perangkat Lunak Untuk Membuat Basis Data Taksonomi Berita
2.	Platform	Dekstop	Dekstop	Web dan android	Web
3.	Kosakata bahasa indonesia	Ya	Ya	Ya	Ya
4.	Tagging	Tidak	Tidak	Ya	Ya
5.	Pengelolaan Melalui Web	Tidak	Tidak	Ya	Ya
6.	Pembuatan basis data kosakata dan kalimat	Tidak	Tidak	Tidak	Ya