

# BAB I Pendahuluan

## 1.1 Latar Belakang

*Tag* merupakan sebuah kata kunci, nama kategori ataupun *metadata*. *Tag* seringkali tidak mengikuti kosakata formal dan dibentuk secara bebas yang dapat merepresentasikan konteks maupun tidak merepresentasikan konteks sama sekali [1]. *Social Tagging* merupakan kegiatan pelabelan (*tagging*) bersama secara publik dan mengkategorisasikan sumber daya yang dapat dibagi (*resources sharing*) pada ruang lingkup daring [2]. Manfaat dari *tagging* juga sebagai alat komunikasi tentang informasi kontekstual dalam sebuah objek. Situs web *social bookmarking* merupakan salah satu contoh dalam pemanfaatan teknologi *social tagging* [3].

Dalam sebuah situs web *social bookmarking* seperti *del.icio.us* terdapat banyak *tag* yang digunakan oleh *user* untuk mengkategorisasikan atau merepresentasikan sebuah situs blog, *URL* dan tautan. *Tag* yang digunakan pada situs *social bookmarking* berguna untuk memudahkan pengguna mencari informasi di masa yang akan datang [2]. Apabila pengguna menggunakan *tag* berulang kali pada setiap situs yang sama maka akan menimbulkan *spam* pada *social tagging*. *Spam* pada *social tagging* juga dapat terjadi apabila pengguna memakai *tag* yang tidak mendeskripsikan konten dari sebuah situs web [4]. Apabila terjadi masalah dalam *tag* seperti *spam*, maka dapat terjadi ambiguitas karena penggunaan *tag* dalam sebuah *bookmark* yang tidak merepresentasikan *bookmark* tersebut. Hal tersebut dapat membuat *user* membutuhkan usaha yang lebih banyak ketika melakukan pencarian pada mesin pencari [3][5].

Dalam tugas akhir ini dikembangkan sebuah aplikasi yang dapat menghitung *spam* sebuah *tag* pada setiap dokumen (*bookmark*) berdasarkan nilai *spam factor*. Penelitian ini menggunakan *dataset* situs *social bookmarking*, *del.icio.us*. *Dataset* yang digunakan berasal dari situs *Hetrec* dan tercantum pada *2nd International Workshop on Information Heterogeneity and Fusion in Recommender Systems (HetRec 2011)* [6]. Pada *dataset* yang didapatkan, hanya digunakan tiga data, yaitu data *user*, *tag*, dan *bookmark*. Data tersebut digunakan untuk membentuk sebuah *posting* [7][8].

Pada penelitian ini, sebuah *tag* mengandung *spam* apabila *tag* tersebut tidak digunakan pada *bookmark* yang benar karena setiap *bookmark* mempunyai daftar *tag* yang benar untuk *bookmark* tersebut [7][8]. Untuk mengetahui sebuah *tag* terindikasi

*spam* atau tidak, akan dibentuk sebuah daftar kumpulan *tag* yang benar untuk setiap dokumen (*bookmark*) yang disebut dengan  $S(d)$  [7][8]. Setelah membentuk  $S(d)$ , maka selanjutnya dibentuk *posting* dari *random good user model* dan *random bad user model*. *Posting* yang telah didapatkan dari kedua model tersebut kemudian digabungkan sehingga membentuk kumpulan *posting*. Selanjutnya digunakan metode *trusted moderator* untuk menghilangkan *posting* yang terindikasi *spam* [7][8]. *Posting* dihilangkan dengan cara mencari *posting* sebanyak fraksi yang telah ditetapkan. Fraksi merupakan persentase *posting* yang dicek secara acak oleh *trusted moderator* [7][8].

Setelah mengimplementasikan *trusted moderator*, kemudian *tag* dan *bookmark* diurutkan berdasarkan jumlah *tag* dan *bookmark* yang paling banyak digunakan pada kumpulan *posting* dengan metode *occurrence-based search*. Untuk perhitungan nilai *spam*, digunakan metode *spam factor* yang menghitung nilai *spam* sebuah *tag* dari kesesuaian penggunaan *tag* tersebut dengan *bookmark* pada setiap *posting* [7][8]. Apabila *tag* tersebut termasuk dalam kumpulan *tag* yang benar pada  $bookmark / S(d)$  maka tidak dianggap sebagai *spam* namun jika tidak termasuk dalam  $S(d)$  ( $T - S(d)$ ) maka dianggap sebagai *spam*. Nilai yang dihasilkan dari *spam factor* berkisar 0 - 1 dimana semakin besar nilai yang didapatkan, maka *tag* semakin terindikasi *spam* [7][8].

## 1.2 Rumusan Masalah

Terkait permasalahan yang dipaparkan, dapat dirumuskan masalah yang diangkat dalam Tugas Akhir ini adalah :

1. Bagaimana cara menghitung nilai *spam* pada sebuah *tag* ?
2. Bagaimana pengaruh *trusted moderator*, *tag budget*, fraksi *trusted moderator*, dan persentase *good user* terhadap nilai *spam factor* yang dihasilkan ?

## 1.3 Tujuan

Terkait rumusan masalah diatas maka dapat ditarik tujuan yang dicapai dalam Tugas Akhir ini, yaitu :

1. Mengimplementasikan metode untuk mengetahui cara menghitung nilai *spam* pada sebuah *tag*.

2. Pengaruh *trusted moderator*, *tag budget*, fraksi *trusted moderator*, dan persentase *good user* terhadap nilai *spam factor* yang dihasilkan.

## 1.4 Batasan Masalah

Batasan Masalah Dalam pengerjaan Tugas Akhir ini terdapat beberapa batasan masalah, yaitu:

1. Daftar *tag* yang benar pada setiap *bookmark / S(d)* pada penelitian ini dibentuk secara acak.
2. *Dataset* yang digunakan merupakan *dataset* yang diunduh dari situs web <http://groupLens.org/datasets/hetrec-2011/> dan hanya menggunakan 3 *file* yaitu *file usertaggedbookmarks-timestamps.csv* untuk *dataset user*, *file tags.csv* untuk *dataset tag*, dan *file bookmarks.csv* untuk *dataset bookmark*.

## 1.5 Metodologi Penyelesaian Masalah

Dalam pengerjaan Tugas Akhir ini terdapat beberapa metodologi penyelesaian masalah, yaitu:

### 1. Identifikasi Masalah

Pada tahap ini diidentifikasi permasalahan yang dipecahkan dalam penelitian ini, seperti membangun set *tag* yang benar ( $S(d)$ ) pada sebuah dokumen (*bookmark*), membangkitkan *posting* dengan cara *random good user model* dan *random bad user model*, menggabungkan *posting random user*, menerapkan metode *trusted moderator*, menerapkan metode *occurrence-based search*, dan menghitung *spam* dengan metode *spam factor*. Setelah masalah ditemukan kemudian menentukan rancangan sistem untuk mengatasi permasalahan tersebut.

### 2. Pengumpulan Data

Pada tahap ini pengumpulan data dilakukan dengan cara mencari *dataset* yang sesuai dengan kebutuhan pada sistem yang dibangun. Data yang dibutuhkan adalah data *user*, *tag* dan dokumen. Untuk data dokumen, karena studi kasus yang diangkat pada tugas akhir ini adalah situs *social bookmarking del.icio.us* maka dokumen dalam tugas akhir ini adalah sebuah *bookmark*.

### 3. Perancangan Sistem

Pada tahap ini dirancang sebuah sistem yang sesuai untuk menyelesaikan permasalahan yang telah diidentifikasi sebelumnya. Sistem yang dibangun

meliputi  $S(d)$ , *random good user model* dan *random bad user model*, *join random user*, *trusted moderator*, *occurrence-based search* dan *spam factor*.

#### 4. Pengujian Sistem

Pada tahap ini sistem yang telah dirancang diuji sesuai dengan kebutuhan dan parameter – parameter yang telah ditetapkan untuk pengujian. Mengubah parameter pada sistem yang telah dibangun, implementasi metode *trusted moderator* pada *posting* yang telah dibangun dari *random good user* dan *random bad user model*, mengurutkan *tag* dengan skema pencarian *occurrence-based search* dan menghitung nilai *spam* pada kumpulan *posting* dengan metode *spam factor*.

#### 5. Analisis Hasil

Untuk analisis maka dibandingkan nilai *spam factor* pada kumpulan *posting* yang mengimplementasikan *trusted moderator* dan kumpulan *posting* yang tidak mengimplementasikan *trusted moderator*. Selain itu, analisis hasil juga membandingkan nilai *spam factor* apabila parameter – parameter pada penelitian diubah.

#### 6. Pembuatan Laporan

Setelah melakukan pengujian sistem dan analisis hasil, selanjutnya adalah mendokumentasikan hasil dari penelitian ini dan melampirkan dokumen pendukung yang berkaitan tentang penelitian ini.

## 1.6 Sistematika Penulisan

Dalam menyelesaikan Tugas Akhir ini, sistematika penulisan yang digunakan adalah sebagai berikut :

#### 1. BAB I : Pendahuluan

Pada Bab ini dijelaskan latar belakang penelitian, masalah yang dihadapi, tujuan penelitian, batasan penelitian, dan metodologi penyelesaian masalah penelitian terkait membangun  $S(d)$  setiap *bookmark*, membangun *posting random user*, penerapan *trusted moderator*, *occurrence-based search* dan *spam factor*.

#### 2. BAB II : Kajian Pustaka

Pada Bab ini dijelaskan dan didefinisikan metode dan istilah yang digunakan, diantaranya adalah *spam*, *social bookmarking*, *tagging system model*, *occurrence-based search*, *trusted moderator* dan *spam factor*.

3. BAB III : Perancangan Sistem

Pada Bab ini dijelaskan gambaran umum sistem dan bagaimana sistem bekerja. Pada bab ini juga dijelaskan setiap tahap pembangunan sistem secara detail, skenario pengujian dan cara perhitungan *spam factor*.

4. BAB IV : Pengujian dan Analisis

Pada Bab ini sistem yang telah dibuat diuji sesuai dengan kebutuhan dan parameter – parameter yang telah ditetapkan untuk menjawab rumusan masalah. Setelah melakukan pengujian maka dilakukan analisis terhadap hasil pengujian yang didapatkan.

5. BAB V : Kesimpulan dan Saran

Pada Bab ini dipaparkan kesimpulan yang didapatkan dari hasil pengujian, selain itu diberikan saran terhadap sistem untuk penelitian yang akan dibuat selanjutnya.