# APLIKASI MESIN PENCARI DOKUMEN *CROSS LANGUAGE* BAHASA INGGRIS – BAHASA INDONESIA MENGGUNAKAN *VECTOR SPACE MODEL*

**SKRIPSI**

**Disusun Sebagai Salah Satu Syarat**
**untuk Memperoleh Gelar Sarjana Komputer**
**pada Jurusan Ilmu Komputer/ Informatika**

**Disusun oleh:**
**ABDUL MAGHFIR ZAKIR**
**24010311130054**

**JURUSAN ILMU KOMPUTER/INFORMATIKA**
**FAKULTAS SAINS DAN MATEMATIKA**
**UNIVERSITAS DIPONEGORO**
**2015**

# HALAMAN PERNYATAAN KEASLIAN SKRIPSI

Saya yang bertanda tangan di bawah ini :

Nama : Abdul Maghfir Zakir

NIM : 24010311130054

Judul : Aplikasi Mesin Pencari Dokumen *Cross Language* Bahasa Inggris – Bahasa Indonesia Menggunakan *Vector Space Model*.

Dengan ini saya menyatakan bahwa dalam tugas akhir/ skripsi ini tidak terdapat karya yang pernah diajukan untuk memperoleh gelar kesarjanaan di suatu Perguruan Tinggi, dan sepanjang pengetahuan saya juga tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang lain, kecuali yang secara tertulis diacu dalam naskah ini dan disebutkan di dalam daftar pustaka.

# HALAMAN PENGESAHAN

Judul : Aplikasi Mesin Pencari Dokumen *Cross Language* Bahasa Inggris – Bahasa Indonesia Menggunakan *Vector Space Model*.

Nama : Abdul Maghfir Zakir

NIM : 24010311130054

Telah diujikan pada sidang tugas akhir tanggal 22 Desember 2015 dan dinyatakan lulus pada tanggal **14 Januari 2016**.

Semarang, Januari 2016

Mengetahui,
Ketua Jurusan Ilmu Komputer/ Informatika
FSM UNDIP


Ragil Saputra, S.Si, M.Cs
NIP. 198010212005011003

Panitia Penguji Tugas Akhir
Ketua,

Drs. Djalal Er Riyanto, M.IKom
NIP. 195412191980031003

# HALAMAN PENGESAHAN

Judul : Aplikasi Mesin Pencari Dokumen *Cross Language* Bahasa Inggris – Bahasa Indonesia Menggunakan *Vector Space Model*.

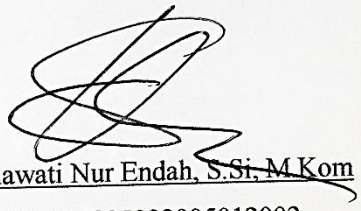
Nama : Abdul Maghfir Zakir

NIM : 24010311130054

Telah diujikan pada sidang tugas akhir tanggal 22 Desember 2015.

Semarang, Januari 2016

Pembimbing



Sukmawati Nur Endah, S.Si, M.Kom

NIP. 197805022005012002

# ABSTRAK

Fenomena banjirnya informasi seiring berkembangnya teknologi dapat ditangani dengan kajian-kajian yang ada pada *information retrieval* (IR). Namun permasalahan pada IR semakin berkembang karena adanya perbedaan bahasa antar negara. Oleh karena itu dibutuhkan suatu aplikasi *cross-language information retrieval* (CLIR) yang mampu mengoptimalkan hasil pencarian dokumen lintas bahasa. Aplikasi mesin pencari dokumen ini dapat melakukan penerjemahan pada masukan *query* lalu dihitung kemiripannya terhadap dokumen yang memiliki bahasa yang berbeda. Nilai kemiripan ini dapat ditentukan berdasarkan nilai similaritas dokumen terhadap *query* dengan *Vector Space Model.* Sepuluh dokumen dengan nilai kemiripan tertinggi akan diambil dan dihitung nilai ketepatannya (*precision*). Berdasarkan hasil evaluasi dengan menggunakan 8 *user* dan 10 *query*, didapatkan nilai *Mean Average Precision* (MAP) keseluruhan sebesar 0.912 atau sekitar 91%.

**Kata kunci** : mesin pencari dokumen, CLIR, *vector space model*

# ABSTRACT

The information flooding phenomenon as technology growth can be handled by existing studies on information retrieval (IR). However, the problem in IR keep on growing because of language differences between countries. Therefore, it needs a cross-language information retrieval (CLIR) application which is able to optimize cross-language document search result. This document search engine application can translate the query input, then calculate its similarity to the document that has a different language. The similarity value can be determined based on the similarity of the document to the query with Vector Space Model. Ten documents with the highest similarity value will be taken and its value of accuracy (precision) will be calculated. Based on the evaluation results by using 8 users and 10 queries, it was discovered that its overall Mean Average Precision (MAP) is 0.912 or about 91%.

**Keyword**: document search engine, CLIR, *vector space model*

# KATA PENGANTAR

Segala puji syukur bagi Tuhan Yang Maha Esa atas karunia-Nya yang diberikan kepada penulis sehingga penulis dapat menyelesaikan penulisan laporan tugas akhir yang berjudul "Aplikasi Mesin Pencari Dokumen *Cross Language* Bahasa Inggris – Bahasa Indonesia Menggunakan *Vector Space Model*". Laporan tugas akhir ini disusun sebagai salah satu syarat untuk memperoleh gelar sarjana strata satu pada Jurusan Ilmu Komputer/ Informatika Fakultas Sains dan Matematika Universitas Diponegoro Semarang.

Dalam penyusunan laporan ini penulis banyak mendapat bimbingan dan bantuan dari berbagai pihak. Untuk itu, pada kesempatan ini penulis mengucapkan rasa hormat dan terima kasih kepada:

1. Ragil Saputra, S.Si, M.Cs selaku Ketua Jurusan Ilmu Komputer/Informatika
2. Helmie Arif Wibawa,S.Si, M.Cs selaku Koordinator Tugas Akhir
3. Sukmawati Nur Endah, S.Si, M.Kom selaku dosen pembimbing
4. Semua pihak yang telah membantu kelancaran dalam penyusunan tugas akhir, yang tidak dapat penulis sebutkan satu persatu.

Penulis menyadari bahwa dalam laporan ini masih banyak kekurangan baik dari segi materi ataupun dalam penyajiannya karena keterbatasan kemampuan dan pengetahuan penulis. Oleh karena itu, kritik dan saran sangat penulis harapkan. Semoga laporan ini dapat bermanfaat bagi pembaca dan penulis pada umumnya.

Semarang,  Januari 2016
Penulis,

Abdul Maghfir Zakir
24010311130054

# DAFTAR ISI

# DAFTAR GAMBAR

# DAFTAR TABEL

# DAFTAR LAMPIRAN

# BAB I
# PENDAHULUAN

## 1.1. Latar Belakang

Jumlah informasi yang tersedia di internet saat ini sangat banyak dan terus bertambah setiap saat. Hal ini dikarenakan perkembangan teknologi yang dimanfaatkan sangat membantu dalam penyebaran informasi. Seiring dengan peningkatan jumlah dan keragaman informasi yang beredar, pengguna semakin sulit mendapatkan informasi yang diinginkan. Kebutuhan penggunapun mulai bergeser dari yang dulunya mencari informasi secara kuantitatif menjadi kualitatif. Salah satu bidang yang membahas tentang pencarian informasi yaitu *information retrieval* (IR). Fenomena banjirnya informasi dapat diselesaikan dengan kajian-kajian yang ada pada IR agar pengguna dapat menemukan informasi yang sesuai keinginan.

Permasalahan pada IR semakin berkembang karena adanya perbedaan bahasa dari berbagai negara. Misalnya dilakukan pencarian informasi dengan memasukkan *query* dengan bahasa tertentu, maka hasil yang diterima pengguna hanya dokumen berisi informasi yang ditulis dengan bahasa tersebut. Sehingga hasil pencarian tidak dapat memberikan hasil yang maksimum untuk pengguna.

Peningkatan hasil pencarian agar lebih optimal pada *search engine* ini dapat diselesaikan dengan *cross-language information retrieval* (CLIR). CLIR pada dasarnya memiliki metode yang sama dengan IR, hanya saja CLIR memiliki tahap penerjemahan pada *query* atau pada dokumen korpus. Kedua metode penerjemahan tersebut memiliki kekurangan masing-masing. Metode dengan menerjemahkan dokumen korpus akan memiliki proses yang lama dan tidak praktis sedangkan untuk menerjemahkan *query* menjadi sulit pada saat *query* bersifat ambigu.

Penerapan CLIR ini terbagi lagi menjadi beberapa model, antara lain *Probabilistic Model*, *Set-theoretic Model* dan *Algebratic Model*. *Probabilistic Model* contohnya pada penerapan *Teorema Bayes*, penerapan *Set-theoretic Models* pada *Standard Boolean* dan *Extended*, dan *Algebratic Model* contohnya adalah *Vector Space Model*. Dari tiga model yang telah disebutkan, *Algebratic Model* dengan contoh *Vector Space Model* adalah model yang paling sederhana dalam pencarian kata. *Vector Space Model* telah terbukti memiliki efektifitas dalam pencarian kata

dengan menampilkan hasil pencariannya berdasar kemiripan *vector query* dan *vector* dokumen (Bari & Saputra, 2011).

Penelitian terdahulu yang menggunakan metode ini antara lain Fatkhul Amin dalam publikasinya yang berjudul Sistem Temu Kembali Informasi dengan Pemeringkatan Metode *Vector Space Model* (Amin, 2013). Penelitian ini membahas tentang mesin pencari dokumen menggunakan *Vector Space Model* dengan hasil evaluasi presisi mencapai 99%. Selain itu dari publikasi Andre Hesel, Ricky Dwiputra dan Hendra yang berjudul Studi dan Evaluasi Kinerja Model-Model *Information Retrieval* Berbasiskan Dokumen Teks menyimpulkan bahwa dalam memilih model *information retrieval* sebaiknya menggunakan *Vector Space Model*. Hal ini dikarenakan pada penelitian diketahui bahwa *Vector Space Model* memberikan *Mean Average Precision* yang lebih besar yaitu 71,25% dibanding dengan *Boolean Model* dan *Latent Semantic Indexing (LSI)* (Hesel, et al., 2013).

Berdasarkan uraian diatas maka akan dibangun sebuah aplikasi mesin pencari dokumen *cross language* Bahasa Inggris – Bahasa Indonesia Menggunakan *vector space model*.

## 1.2. Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, maka dapat dirumuskan masalah yaitu bagaimana membuat aplikasi mesin pencari dokumen *cross language* Bahasa Inggris – Bahasa Indonesia Menggunakan *vector space model*.

## 1.3. Tujuan dan Manfaat

Penelitian Tugas Akhir ini bertujuan untuk menghasilkan suatu aplikasi yang dapat mencari dokumen dengan kemampuan *cross-language* Bahasa Inggris – Bahasa Indonesia Menggunakan *vector space model*.

Aplikasi mesin pencari ini diharapkan bisa bermanfaat dalam menemukan dokumen yang relevan dengan Menggunakan *query* Bahasa Inggris untuk menemukan dokumen yang relevan dalam Bahasa Indonesia sekalipun, begitupun sebaliknya.

## 1.4. Ruang Lingkup

Ruang lingkup dari pengembangan Aplikasi Mesin Pencari Dokumen *Cross Language* Bahasa Inggris – Bahasa Indonesia Menggunakan *Vector Space Model* adalah sebagai berikut:

1. Penerapan CLIR yang digunakan adalah penerjemahan pada *query*.
2. Bahasa yang digunakan adalah Bahasa Inggris dan Bahasa Indonesia.
3. Dokumen diolah berasal dari korpus lokal, bukan menggunakan *web crawling*.
4. Dokumen yang diolah menggunakan korpus menggunakan format file txt.
5. Dokumen yang ter-*retrieve* dapat diunduh dengan format pdf.
6. Evaluasi metode menggunakan minimal 20 dokumen jurnal Bahasa Indonesia dan 20 dokumen jurnal Bahasa Inggris.
7. Pengembangan aplikasi menggunakan model proses pengembangan perangkat lunak waterfall.
8. Pembangunan aplikasi menggunakan Bahasa Pemrograman PHP dan *Database Management System* MySQL.

## 1.5. Sistematika Penulisan

Sistematika penulisan yang digunakan dalam tugas akhir ini dibagi menjadi dalam beberapa pokok bahasan, yaitu:

BAB I PENDAHULUAN

Bab ini membahas latar belakang, rumusan masalah, tujuan dan manfaat, ruang lingkup dan sistematika penulisan dalam pembuatan tugas akhir megenai Aplikasi Mesin Pencari Dokumen *Cross-Language* Bahasa Inggris – Bahasa Indonesia Menggunakan *Vector Space Model*.

BAB II TINJAUAN PUSTAKA

Bab ini menyajikan tinjauan pustaka yang berhubungan dengan topik tugas akhir. Tinjauan pustaka yang digunakan dalam penyusunan tugas akhir ini meliputi *information retrieval*, arsitektur *information retrieval*, proses pengembangan perangkat lunak, PHP, MySQL, *Vector Space Model*, dan evaluasi.

BAB III ANALISIS DAN PERANCANGAN

Bab ini menyajikan mengenai pembahasan tahapan dari model pengembangan perangkat lunak waterfall yang meliputi tahap analisis dan perancangan dari Aplikasi Mesin Pencari Dokumen *Cross-Language* Bahasa Inggris – Bahasa Indonesia Menggunakan *Vector Space Model*.

BAB IV IMPLEMENTASI DAN PENGUJIAN

Bab ini menyajikan mengenai pembahasan tahapan dari model pengembangan perangkat lunak waterfall yang meliputi tahap implementasi dan pengujian dari Aplikasi Mesin Pencari Dokumen *Cross-Language* Bahasa Inggris – Bahasa Indonesia Menggunakan *Vector Space Model*.

BAB V PENUTUP

Bab ini merupakan kesimpulan dari bab-bab yang telah dibahas sebelumya serta saran untuk pengembangan penelitian tugas akhir lebih lanjut.