Using query transformation to improve Gnutella search performance

Surendar Chandra (surendar@acm.org)

William Acosta (william.acosta@utoledo.edu)

attempt at matching shared filenames and queries to improve search performance

Role of Gnutella filenames and queries

- Gnutella query resolution poor: ~10% success
  - overlay and search improvements can help
- shared filenames and queries are uncoordinated
- all query terms must match shared filename terms
- consider query Q:$< q_1 \; q_2 \; q_3 >$
  - matches F:$< q_3 \; q_1 \; f_1 \; q_2 >$
  - does not match F:$< q_1 \; q_2 \; q_3' >$, F:$< f_1 \; f_2 \; f_3 >$ ….

Empirical analysis

- shared filenames from crawler:
  - April 2007: 20 million files, 37 thousand peers
  - Feb 2008: 17 million files, 34 thousand peers
- queries from instrumented Gnutella client
- ~56% queries had no matching objects
  - overlay agnostic analysis

- **Understanding the Practical Limits of the Gnutella P2P System: An Analysis of Query Terms and Object Name Distributions**, *William Acosta* and *Surendar Chandra*. In MMCN ’08, Jan ’08

Approach: transform queries to match filenames

- intuition: queries are inherently related to shared filenames: more F:< q_1 q_2 q_3' > than F:< f_1 f_2 f_3 >
- Challenges:
  - identifying files related to the intent of the original query
    - only choose keywords from original query
  - limiting scope
    - intuition: inappropriate transformations will match more files than typical
      - typical match - 25 files. match any keyword > 24K
  - practical
    - use information from neighbors

Transformations investigated

- correct misspelt keywords: Q:< q_1 q_2 q_3' >
  - unlike Zaharia, used file terms from peer neighborhood
- remove keywords:Q:< q_1 q_2 >, Q:< q_1 >
  - tried queries of length one, two and three
  - policies:
    - random: randomly drop keywords
    - popular: choose popular terms from peer neighborhood
    - co-popular: co-occurrence popularity of pairs of keywords
    - hybrid: spell+co-popular

Spell

- 30% of failed queries matched 25 files
- improvement over dictionary based approach:
  - 17% of queries different character-set (more multi-lingual)
  - many song names use slangs (e.g. Dat)
  - terms change with release of new songs

Removing keywords

- random:
  - 26%, 32%, 39% of failed queries: failed (remove 1, 2 or 3 terms)
  - 52%, 55%, 57% of transformed queries matched < 25 files
- popular:
  - 21%, 35%, 44% of failed queries: failed
  - 45%, 54%, 58% of transformed queries matched < 25 files
- co-popular:
  - 17%, 30%, 46% of failed queries: failed
  - 39%, 47%, 56% of transformed queries matched < 25 files

Hybrid approach

  - 16%, 28%, 44% of failed queries: fail
  - 43%, 47%, 55% of transformed queries < 25 files
- choosing 3 keywords, success rate from 45% to
  - 73% - spell
  - 79% - random
  - 76% - popular
  - 75% - co-popular
  - 76% - hybrid

Peer neighborhood size

- tried neighborhood size of 64, 200 and 400
  - randomly picked peer neighbors
  - results robust and so use 64

Middleware

- operate as ultra-peer, collect information about leaf peers during handshake
  - compute co-occurrence and popularity
- issue original and transformed query
  - original query succeeds - discard transformed query
- ignore bogus peers - some peers always succeed
- subjectively - 61% of failed queries succeed
  - query issuer’s intent not always clear

Subjective results - success

- original: “**barbara streisen woman love**”
  - transform: “barbara woman love”
    - barbara streisand - woman in love.mp3
    - barbara streisand & beegees - wild flower - woman in love.mp3
    - barbara striesand - i am a woman in love.mp3
    - Bee Gees & Barbara Streisand - Woman In Love.mp3
    - Barbara Streisand - I am a Woman In Love.mp3

Subjective results - failure

- original “**o dublado retorno superman**”
- transform: “superman”
  - Soulja Boy - superman dat hoe.mp3
  - MTV MashUps - Eminem vs Justin Timberlake - Cry my a superman.mp3
  - Dave Matthews Band - Superman.mp3
  - Souljah boy ft. Twista- Crank Dat Superman (Remix).mp3
  - Lyfe Jennings - The Phoenix - 06 - Ghetto Superman.mp3
  - Superman Returns 720p HD DVDRip x264 DD5 1-HINT.zip
  - Coldplay - Superman.mp3

Subjective result count

- original: “**snoop dogg feb concert bercy hipnotize game france live**”
- transform: ”snoop dogg game”
  - 91 results

- original: “**boy walking out of stride zip**”
- transform: “boy walking out”
  - 1 result

Summary

- Gnutella queries fail because of mis-match in queries and filenames
- investigated practical ways to transform query
  - defined notion of relevance to intent of original query
  - success rates up from 44% to ~75%
- middleware
  - subjective analysis: ~60% success for failed queries (~74%)