Semantic Search for Quantity Expressions

Tom Wiesing

Supervisor: Michael Kohlhase
Co-supervisor: Tobias Preusser

May 20, 2015
110392 Guided Research Applied and Computational Mathematics & Thesis

Overview

Overview

- Motivation: Problem and State Of The Art

Overview

- Motivation: Problem and State Of The Art
- Our Approach: Structure Of The Search Engine

Overview

- Motivation: Problem and State Of The Art
- Our Approach: Structure Of The Search Engine
  - The Unit System

Overview

- Motivation: Problem and State Of The Art
- Our Approach: Structure Of The Search Engine
  - The Unit System
  - The Search Algorithm

Overview

- Motivation: Problem and State Of The Art
- Our Approach: Structure Of The Search Engine
  - The Unit System
  - The Search Algorithm
  - The Frontend

Overview

- Motivation: Problem and State Of The Art
- Our Approach: Structure Of The Search Engine
  - The Unit System
  - The Search Algorithm
  - The Frontend
- The Implementation

Overview

- Motivation: Problem and State Of The Art
- Our Approach: Structure Of The Search Engine
  - The Unit System
  - The Search Algorithm
  - The Frontend
- The Implementation
- Time for Questions

Motivation (1)

- We use units every day

Motivation (1)

- We use units every day
- We encounter them everywhere:

Motivation (1)

- We use units every day
- We encounter them everywhere:
  - When driving, there are speed limits, for example: 

$\frac{\text{km}}{\text{h}}$

Motivation (1)

- We use units every day
- We encounter them everywhere:
  - When driving, there are speed limits, for example: 

$\frac{\text{km}}{\text{h}}$
  - When baking, it often says in recepies something like: “add 3 tea spoons of sugar”

Motivation (1)

- We use units every day
- We encounter them everywhere:
  - When driving, there are speed limits, for example: 

$\frac{\text{km}}{\text{h}}$
  - When baking, it often says in recepies something like: "add 3 tea spoons of sugar"
  - When shopping for shoes there are different sizes

Motivation (1)

- We use units every day
- We encounter them everywhere:
  - When driving, there are speed limits, for example: 

$\frac{\text{km}}{\text{h}}$
  - When baking, it often says in recepies something like: "add 3 tea spoons of sugar"
  - When shopping for shoes there are different sizes
- In scientific papers they occur a lot

Motivation (1)

- We use units every day
- We encounter them everywhere:
  - When driving, there are speed limits, for example: 

$\frac{\text{km}}{\text{h}}$
  - When baking, it often says in recepies something like: "add 3 tea spoons of sugar"
  - When shopping for shoes there are different sizes
- In scientific papers they occur a lot
- everything which somehow models a real system has at least one quantity expression

Motivation (1)

- We use units every day
- We encounter them everywhere:
  - When driving, there are speed limits, for example: 

$\frac{\text{km}}{\text{h}}$
  - When baking, it often says in recepies something like: "add 3 tea spoons of sugar"
  - When shopping for shoes there are different sizes
- In scientific papers they occur a lot
- everything which somehow models a real system has at least one quantity expression
- everything is quantified

Motivation (2)

- But where is the problem?

Motivation (2)

- But where is the problem?
- Within one paper only a handful of units is used

Motivation (2)

- But where is the problem?
- Within one paper only a handful of units is used
- In general there are **a lot** of **different** units to describe **the same** quantity

Motivation (2)

- But where is the problem?
- Within one paper only a handful of units is used
- In general there are **a lot** of **different** units to describe **the same** quantity
- Just for lengths:

Motivation (2)

- But where is the problem?
- Within one paper only a handful of units is used
- In general there are **a lot** of **different** units to describe **the same** quantity
- Just for lengths:

Motivation (2)

- But where is the problem?
- Within one paper only a handful of units is used
- In general there are **a lot** of **different** units to describe **the same** quantity
- Just for lengths: *Meter*,

Motivation (2)

- But where is the problem?
- Within one paper only a handful of units is used
- In general there are **a lot** of **different** units to describe **the same** quantity
- Just for lengths: *Meter*, *Inch*,

Motivation (2)

- But where is the problem?
- Within one paper only a handful of units is used
- In general there are **a lot** of **different** units to describe **the same** quantity
- Just for lengths: *Meter*, *Inch*, *Foot*,

Motivation (2)

- But where is the problem?
- Within one paper only a handful of units is used
- In general there are **a lot** of **different** units to describe **the same** quantity
- Just for lengths: *Meter*, *Inch*, *Foot*, *Mile*,

Motivation (2)

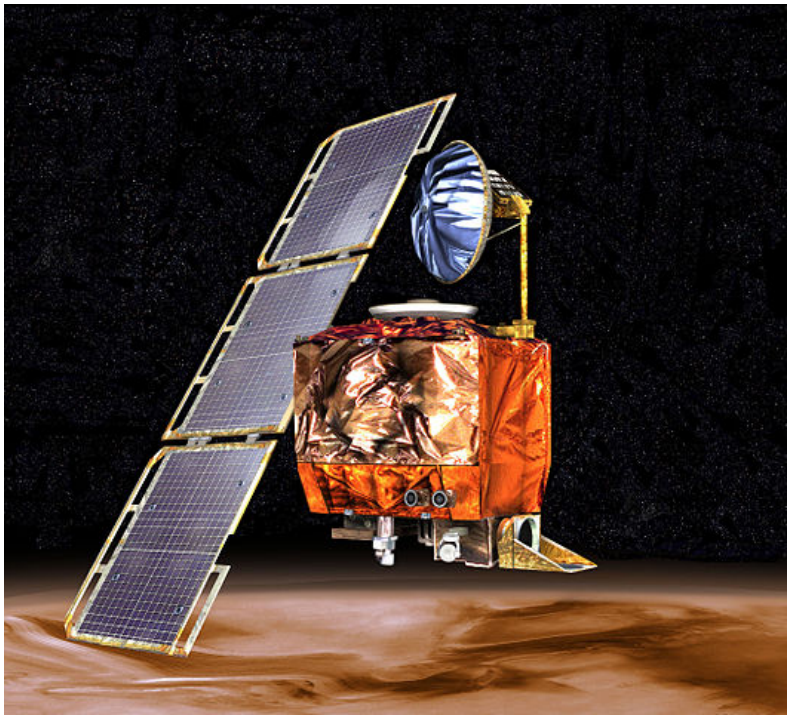
- But where is the problem?
- Within one paper only a handful of units is used
- In general there are **a lot** of **different** units to describe **the same** quantity
- Just for lengths: *Meter*, *Inch*, *Foot*, *Mile*, *Nautical Mile*,

Motivation (2)

- But where is the problem?
- Within one paper only a handful of units is used
- In general there are **a lot** of **different** units to describe **the same** quantity
- Just for lengths: *Meter*, *Inch*, *Foot*, *Mile*, *Nautical Mile*, ...
- This can cause problems when not converting properly

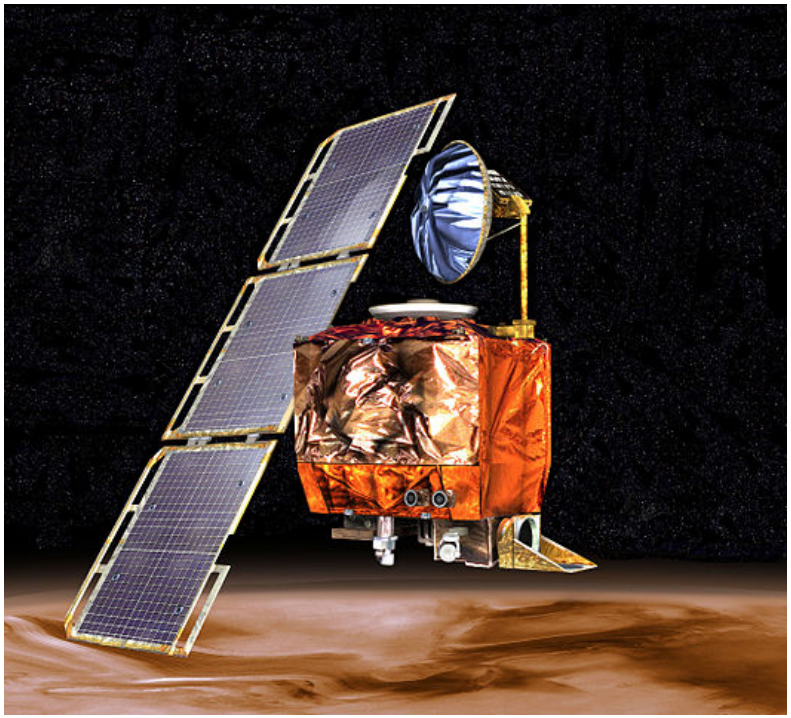
Motivation (2)

- But where is the problem?
- Within one paper only a handful of units is used
- In general there are **a lot** of **different** units to describe **the same** quantity
- Just for lengths: *Meter*, *Inch*, *Foot*, *Mile*, *Nautical Mile*, ...
- This can cause problems when not converting properly
- Mars Climate Orbiter (1999)



Motivation (2)

- But where is the problem?
- Within one paper only a handful of units is used
- In general there are **a lot** of **different** units to describe **the same** quantity
- Just for lengths: *Meter*, *Inch*, *Foot*, *Mile*, *Nautical Mile*, ...
- This can cause problems when not converting properly
- Mars Climate Orbiter (1999)



- Different Units are a big problem

Motivation (3)

Motivation (3)

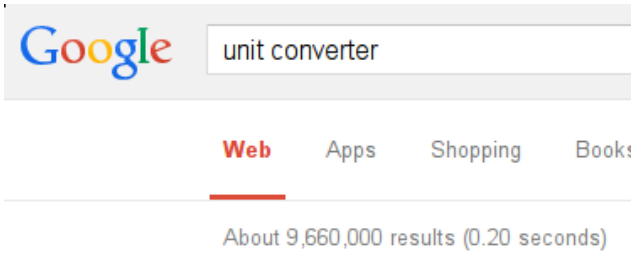
- What is the most common solution?

Motivation (3)

- What is the most common solution?
- Unit Converters

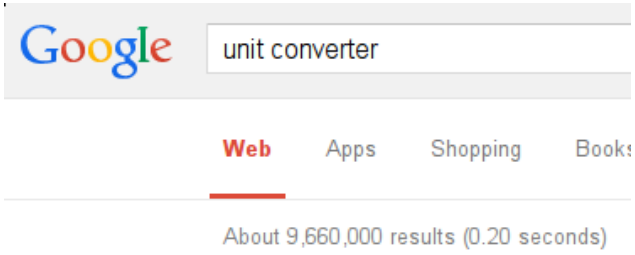
Motivation (3)

- What is the most common solution?
- Unit Converters
  - There are a lot of these

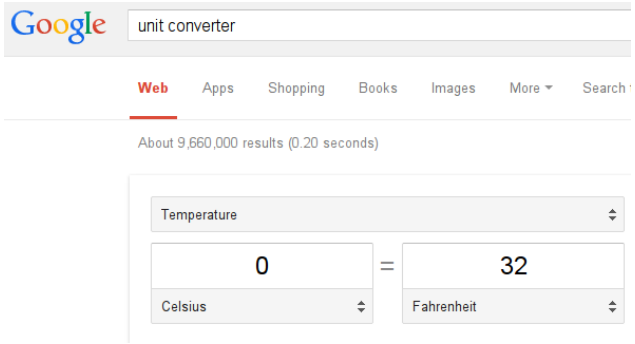


Motivation (3)

- What is the most common solution?
- Unit Converters
  - There are a lot of these



  - Google itself has one integrated



Motivation (4)

- user needs to find out that conversion is required

Motivation (4)

- user needs to find out that conversion is required
- both input **and** output units need to be **given**

Motivation (4)

- user needs to find out that conversion is required
- both input **and** output units need to be **given**
- this is not integrated into the search process itself

Motivation (4)

- user needs to find out that conversion is required
- both input **and** output units need to be **given**
- this is not integrated into the search process itself
- Wouldn't it be nice:

Motivation (4)

- user needs to find out that conversion is required
- both input **and** output units need to be **given**
- this is not integrated into the search process itself
- Wouldn't it be nice:
  - when searching for 25 $\frac{\mathrm{m}}{\mathrm{s}}$

Motivation (4)

- user needs to find out that conversion is required
- both input **and** output units need to be **given**
- this is not integrated into the search process itself
- Wouldn't it be nice:
  - when searching for 25 $\frac{\mathrm{m}}{\mathrm{s}}$
  - we also find 90 $\frac{\mathrm{km}}{\mathrm{h}}$

Motivation (4)

- user needs to find out that conversion is required
- both input **and** output units need to be **given**
- this is not integrated into the search process itself
- Wouldn't it be nice:
  - when searching for 25 $\frac{\mathrm{m}}{\mathrm{s}}$
  - we also find 90 $\frac{\mathrm{km}}{\mathrm{h}}$
  - we did not have to search for all the representations of the same quantity expression

Motivation (4)

- user needs to find out that conversion is required
- both input **and** output units need to be **given**
- this is not integrated into the search process itself
- Wouldn't it be nice:
  - when searching for 25 $\frac{\mathrm{m}}{\mathrm{s}}$
  - we also find 90 $\frac{\mathrm{km}}{\mathrm{h}}$
  - we did not have to search for all the representations of the same quantity expression
- This is the kind of search engine we have built

Our Approach (1)

- What components do we need for a semantic search engine?

Our Approach (1)

- What components do we need for a semantic search engine?
  1. A *Unit System* that is aware of the different representations of a QE

Our Approach (1)

- What components do we need for a semantic search engine?
  1. A *Unit System* that is aware of the different representations of a QE
  2. A *Spotter* that finds representations of QEs inside documents

Our Approach (1)

- What components do we need for a semantic search engine?
  1. A *Unit System* that is aware of the different representations of a QE
  2. A *Spotter* that finds representations of QEs inside documents
  3. A *Search Algorithm* that given a QE finds all its representations in the system

Our Approach (1)

- What components do we need for a semantic search engine?
  1. A *Unit System* that is aware of the different representations of a QE
  2. A *Spotter* that finds representations of QEs inside documents
  3. A *Search Algorithm* that given a QE finds all its representations in the system
  4. A *Frontend* that allows queries to be made

Our Approach (1)

- What components do we need for a semantic search engine?
  1. A *Unit System* that is aware of the different representations of a QE
  2. A *Spotter* that finds representations of QEs inside documents
  3. A *Search Algorithm* that given a QE finds all its representations in the system
  4. A *Frontend* that allows queries to be made
- Spotter is done by *Stiv Sherko*

Our Approach (1)

- What components do we need for a semantic search engine?
  1. A *Unit System* that is aware of the different representations of a QE
  2. A *Spotter* that finds representations of QEs inside documents
  3. A *Search Algorithm* that given a QE finds all its representations in the system
  4. A *Frontend* that allows queries to be made
- Spotter is done by *Stiv Sherko*
- We only need to worry about the *Unit System*, the *Search Algorithm* and the *Frontend*

Our Approach: The Unit System (1)

- Let's start with the Unit System

Our Approach: The Unit System (1)

- Let's start with the Unit System
- We orient ourselves on the SI specification

Our Approach: The Unit System (1)

- Let's start with the Unit System
- We orient ourselves on the SI specification
- Each quantity has a dimension

Our Approach: The Unit System (1)

- Let's start with the Unit System
- We orient ourselves on the SI specification
- Each quantity has a dimension
- According to SI there are 7 basic ones:

Our Approach: The Unit System (1)

- Let's start with the Unit System
- We orient ourselves on the SI specification
- Each quantity has a dimension
- According to SI there are 7 basic ones:
  - length

Our Approach: The Unit System (1)

- Let's start with the Unit System
- We orient ourselves on the SI specification
- Each quantity has a dimension
- According to SI there are 7 basic ones:
  - length
  - mass

Our Approach: The Unit System (1)

- Let's start with the Unit System
- We orient ourselves on the SI specification
- Each quantity has a dimension
- According to SI there are 7 basic ones:
  - length
  - mass
  - time

Our Approach: The Unit System (1)

- Let's start with the Unit System
- We orient ourselves on the SI specification
- Each quantity has a dimension
- According to SI there are 7 basic ones:
  - length
  - mass
  - time
  - electric current

Our Approach: The Unit System (1)

- Let's start with the Unit System
- We orient ourselves on the SI specification
- Each quantity has a dimension
- According to SI there are 7 basic ones:
  - length
  - mass
  - time
  - electric current
  - temperature

Our Approach: The Unit System (1)

- Let's start with the Unit System
- We orient ourselves on the SI specification
- Each quantity has a dimension
- According to SI there are 7 basic ones:
  - length
  - mass
  - time
  - electric current
  - temperature
  - luminous intensity

Our Approach: The Unit System (1)

- Let's start with the Unit System
- We orient ourselves on the SI specification
- Each quantity has a dimension
- According to SI there are 7 basic ones:
  - length
  - mass
  - time
  - electric current
  - temperature
  - luminous intensity
  - amount of substance

Our Approach: The Unit System (1)

- Let's start with the Unit System
- We orient ourselves on the SI specification
- Each quantity has a dimension
- According to SI there are 7 basic ones:
  - length
  - mass
  - time
  - electric current
  - temperature
  - luminous intensity
  - amount of substance
- but there are also quantities where we just *count*

Our Approach: The Unit System (1)

- Let's start with the Unit System
- We orient ourselves on the SI specification
- Each quantity has a dimension
- According to SI there are 7 basic ones:
  - length
  - mass
  - time
  - electric current
  - temperature
  - luminous intensity
  - amount of substance
- but there are also quantities where we just *count*
- and *dimensionless quantities* (such as Information)

Our Approach: The Unit System (1)

- Let's start with the Unit System
- We orient ourselves on the SI specification
- Each quantity has a dimension
- According to SI there are 7 basic ones:
  - length
  - mass
  - time
  - electric current
  - temperature
  - luminous intensity
  - amount of substance
- but there are also quantities where we just *count*
- and *dimensionless quantities* (such as Information)
- so we have 9 basic dimensions

Our Approach: The Unit System (2)

- we can also multiply these to get new dimensions

Our Approach: The Unit System (2)

- we can also multiply these to get new dimensions
- such as area $=$ length $\cdot$ length

Our Approach: The Unit System (2)

- we can also multiply these to get new dimensions
- such as area $=$ length $\cdot$ length
- similarly we can divide dimensions

Our Approach: The Unit System (2)

- we can also multiply these to get new dimensions
- such as area $=$ length $\cdot$ length
- similarly we can divide dimensions
- We can use this to get a meta-mathematical formalisation of dimensions

Our Approach: The Unit System (3)

Dimension		
dimension	:	type
none	:	dimension
count	:	dimension
length	:	dimension
mass	:	dimension
time	:	dimension
current	:	dimension
temperature	:	dimension
luminous	:	dimension
amount	:	dimension
$\cdot$	:	dimension $\rightarrow$ dimension $\rightarrow$ dimension
$\backslash$	:	dimension $\rightarrow$ dimension $\rightarrow$ dimension

Our Approach: The Unit System (4)

- how can we formalise a quantity expressions?

Our Approach: The Unit System (4)

- how can we formalise a quantity expressions?
- In our model QEs can be one of the following:

Our Approach: The Unit System (4)

- how can we formalise a quantity expressions?
- In our model QEs can be one of the following:
  1. (1 times) a *primitive unit*, such as Meter

Our Approach: The Unit System (4)

- how can we formalise a quantity expressions?
- In our model QEs can be one of the following:
  1. (1 times) a *primitive unit*, such as Meter
  2. A *multiplication* of a (real) number with an existing QE, such as 5 Meter

Our Approach: The Unit System (4)

- how can we formalise a quantity expressions?
- In our model QEs can be one of the following:
  1. (1 times) a *primitive unit*, such as Meter
  2. A *multiplication* of a (real) number with an existing QE, such as 5 Meter
  3. A *division* of an existing QE by a (non-zero real) number (equivalent to the above)

Our Approach: The Unit System (4)

- how can we formalise a quantity expressions?
- In our model QEs can be one of the following:
  1. (1 times) a *primitive unit*, such as Meter
  2. A *multiplication* of a (real) number with an existing QE, such as 5 Meter
  3. A *division* of an existing QE by a (non-zero real) number (equivalent to the above)
  4. The *product* of two existing QEs such as Newton $\cdot$ Second

Our Approach: The Unit System (4)

- how can we formalise a quantity expressions?
- In our model QEs can be one of the following:
  1. (1 times) a *primitive unit*, such as Meter
  2. A *multiplication* of a (real) number with an existing QE, such as 5 Meter
  3. A *division* of an existing QE by a (non-zero real) number (equivalent to the above)
  4. The *product* of two existing QEs such as Newton $\cdot$ Second
  5. The *sum* of two existing QEs (of the same dimension)

Our Approach: The Unit System (4)

- how can we formalise a quantity expressions?
- In our model QEs can be one of the following:
  1. (1 times) a *primitive unit*, such as Meter
  2. A *multiplication* of a (real) number with an existing QE, such as 5 Meter
  3. A *division* of an existing QE by a (non-zero real) number (equivalent to the above)
  4. The *product* of two existing QEs such as Newton $\cdot$ Second
  5. The *sum* of two existing QEs (of the same dimension)
  6. The *quotient* of two existing QEs such as 1 $\frac{\text{Meter}}{\text{Second}}$

Our Approach: The Unit System (4)

- how can we formalise a quantity expressions?
- In our model QEs can be one of the following:
  1. (1 times) a *primitive unit*, such as Meter
  2. A *multiplication* of a (real) number with an existing QE, such as 5 Meter
  3. A *division* of an existing QE by a (non-zero real) number (equivalent to the above)
  4. The *product* of two existing QEs such as Newton $\cdot$ Second
  5. The *sum* of two existing QEs (of the same dimension)
  6. The *quotient* of two existing QEs such as 1 $\frac{\text{Meter}}{\text{Second}}$
- this results in the following formalisation:

Our Approach: The Unit System (5)

Quantity Expression		
QE	:	$\text{dimension} \rightarrow \text{type}$
QENMul	:	$\forall x : \text{dimension.number} \rightarrow \text{QE}\,(x) \rightarrow \text{QE}\,(x)$
QENDiv	:	$\forall x : \text{dimension.QE}\,(x) \rightarrow \text{number} \rightarrow \text{QE}\,(x)$
QEMul	:	$\forall x : \text{dimension.}\forall y : \text{dimension.QE}\,(x) \rightarrow \text{QE}\,(y$
QEAdd	:	$\forall x : \text{dimension.QE}\,(x) \rightarrow \text{QE}\,(x) \rightarrow \text{QE}\,(x)$
QEDiv	:	$\forall x : \text{dimension.}\forall y : \text{dimension.QE}\,(x) \rightarrow \text{QE}\,(y$